TUHH
Hamburg
University of
Technology

MUR
Ministero dell'Università e della Ricerca

SERICS
SECURITY AND RIGHTS IN THE CYBERSPACE

Journal First @ ACM Transactions on Software Engineering

Early and Realistic Exploitability Prediction of Just-Disclosed Software Vulnerabilities

How Reliable Can It Be?

Emanuele Iannone, Giulia Sellitto, Emanuele Iaccarino,
Filomena Ferrucci, Andrea De Lucia, Fabio Palomba

emanuele.iannone@tuhh.de

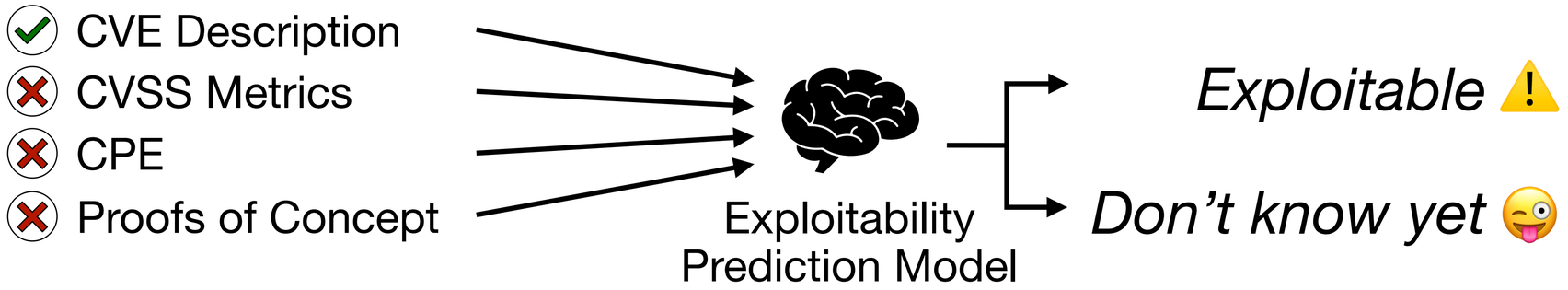
https://emaiannone.github.io/

@EmanueleIannon3

Preprint

Exploitability Prediction

Exploitability prediction: *"Is the given vulnerability going to be exploited soon?"*



Unfortunately, **"just-disclosed"** vulnerabilities do not have such information!

For example, the CVSS assessment is not available before

2 months

average of vulnerabilities before 2021

Preprint

Study Objective

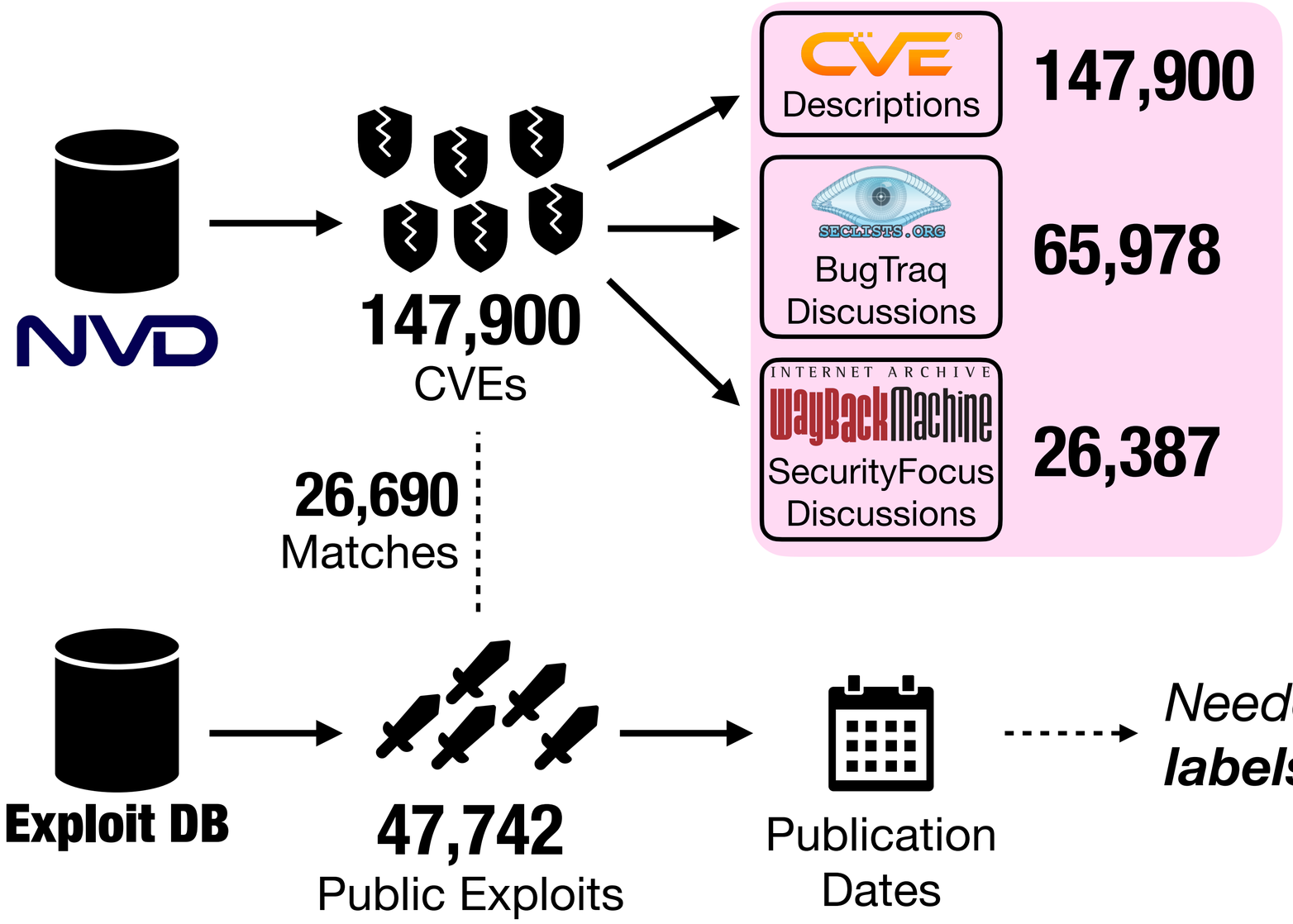
Early prediction uses only what is available at a given time (**early data**).

Goal: Assess exploitability prediction models trained solely on **early data**.

- 👉 **Partially considered** in existing research (two papers at the time of this study).
- 👉 Establish a **baseline** for early prediction rather than having the most effective model.

Preprint

Data Collection

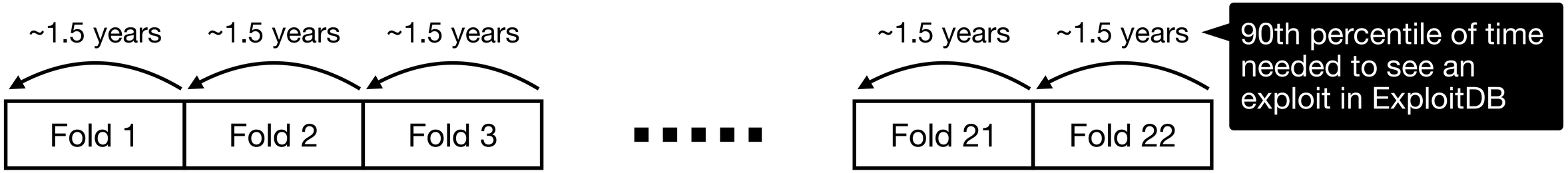


Preprint

Evaluation Setting

We adopted a **time-aware** and **incremental** evaluation setting:

- **Time-aware**: Sort vulnerabilities by the CVE disclosure date.
- **Incremental**: Split into **22 folds** (with ~1.5 years of vulnerabilities), used incrementally. The first 80% is for training, the remaining 20% for testing.



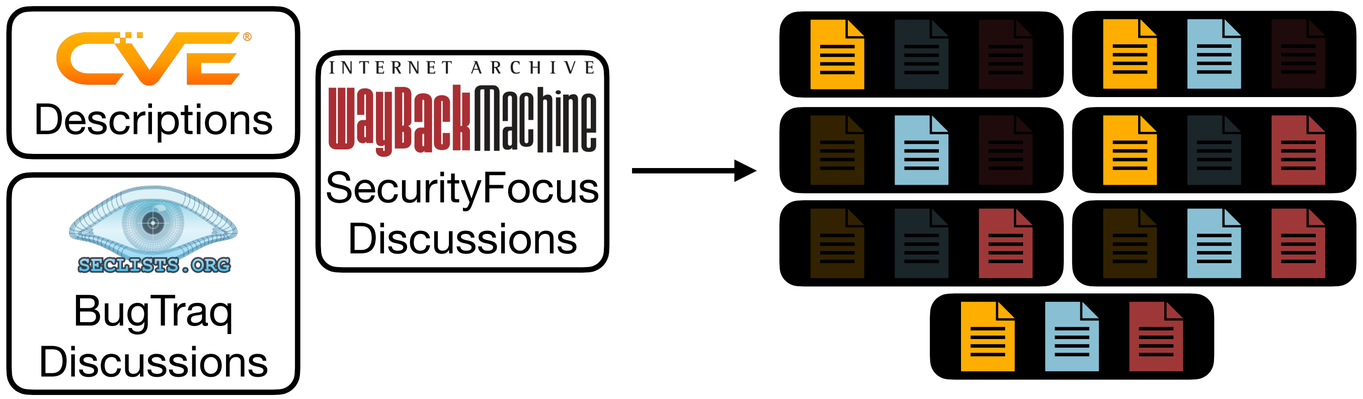
Realistic labeling

- Training instances are labeled as “Exploitable” only if an exploit occurred **before the date of the last training instance**.
- Testing instances are labeled as “Exploitable” only if an exploit occurred **before the date of the last testing instance**.

Metrics: F1 and MCC (also weighted on the training set size).

Preprint

Experimented Configurations



7
Text Corpora

RQ1 Impact of **early data source combinations**

4 input representations (text to features)
- 👉 Bag-of-Words, Term Frequency, TF-IDF, Doc2Vec

3 training data balancing schemas
- 👉 SMOTE, NearMiss (v3), None

6 learning algorithms
- 👉 Random Forest, Decision Tree, SVM, Logistic Regression, Naive Bayes, KNN

RQ2 Impact of **learning configurations**

72
Learning Configurations

Preprint

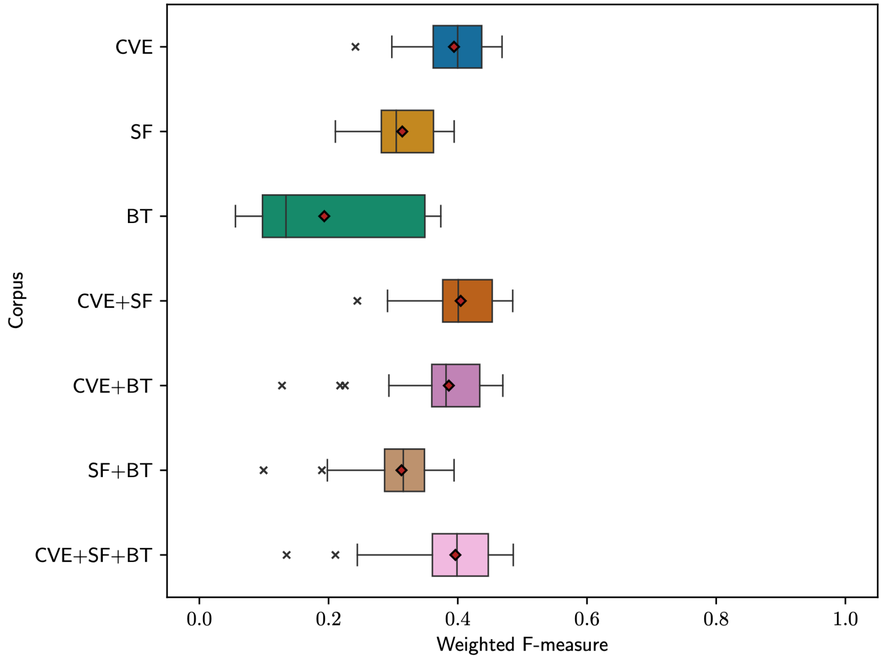
RQ1: Impact of Data Sources

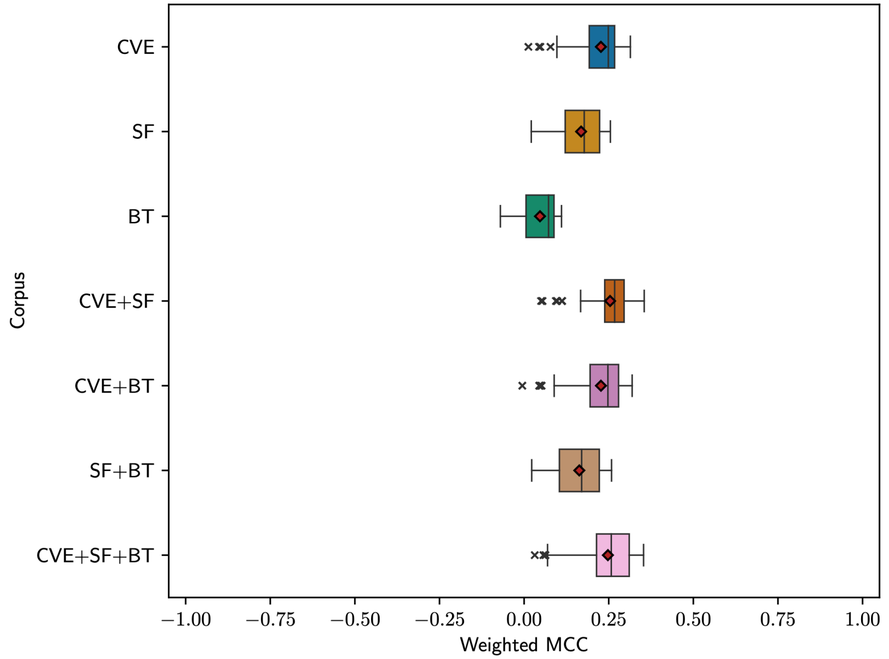
Best Corpus

CVE + SecurityFocus is the best overall.

⚠ No model scored an average of more than:

- **0.48 F1 Score**
- **0.35 MCC**



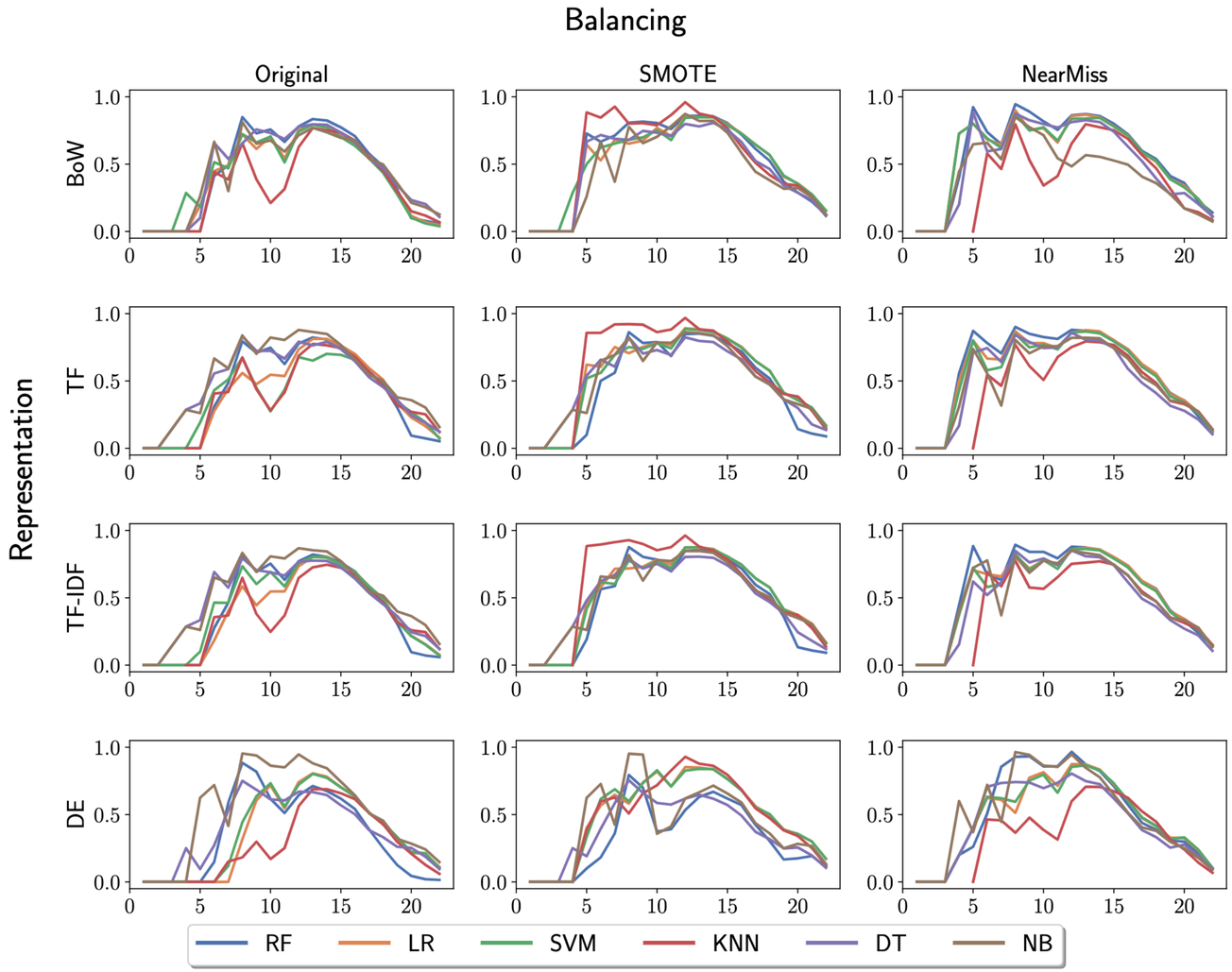


CVE descriptions are enough!

- CVE descriptions are the **best** for early predictions.
- The results are even **worse without them**.
- Online discussions contribute **marginally**, but they do not harm either.

Preprint

RQ2: Impact of Learning Configurations



Zoom: ***CVE + SecurityFocus*** *corpus (F1 score).*

Best configuration

- Weighting with **Term Frequency**
- **SMOTE** oversampling
- **Logistic Regression** model

All models share a similar trend

- Very good in central rounds (8th-15th), with peak on the **15th** (up to 2011).
- Drop toward zero in final rounds (after 15th)!!

In central rounds, F1 was 0.8+, but MCC was around 0.0!
The test sets had almost exclusively true instances!

Preprint

Emanuele Iannone, *Giulia Sellitto, Emanuele Iaccarino, Filomena Ferrucci, Andrea De Lucia, Fabio Palomba*

The Road Ahead

Are more recent vulnerabilities more difficult to exploit?

Difficult to prove, but it does not seem the case. Maybe people are using ExploitDB less…

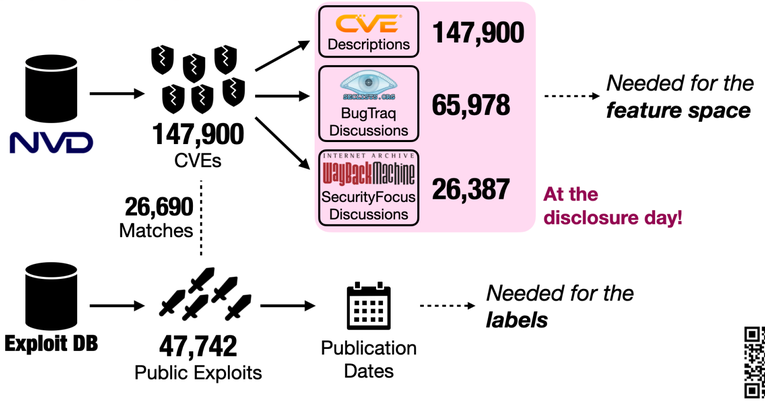
Is ExploitDB a reliable source for the ground truth?

- ☛ We observe that the number of vulnerabilities in 2010-2020 is **double** the number in 2000-2010. But the number of exploits published was **halved**!
- ☛ Aggregating other exploit catalogs does not help much (**too small**).

Call to action: Enhance the **tracking of exploited vulnerabilities in the wild** for a more reliable ground truth.

Preprint

Data Collection



Emanuele Iannone, Giulia Sellitto, Emanuele Iaccarino, Filomena Ferrucci, Andrea De Lucia, Fabio Palomba

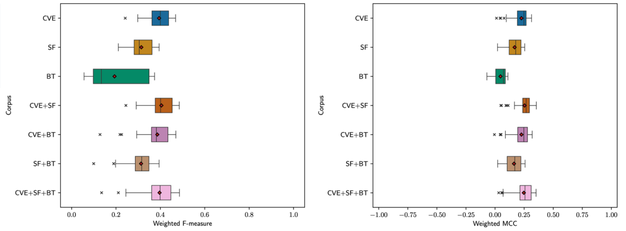
Preprint

Early and Realistic Exploitability Prediction of Just-Disclosed Software Vulnerabilities

RQ1: Impact of Data Sources

Best Corpus
CVE + SecurityFocus is the best overall.

⚠ No model scored an average of more than:
- **0.48 F1 Score**
- **0.35 MCC**



BugTraq Discussions
- Much **worse** than other corpora (Nemenyi tests).
- Give **no contribution** to other corpora.

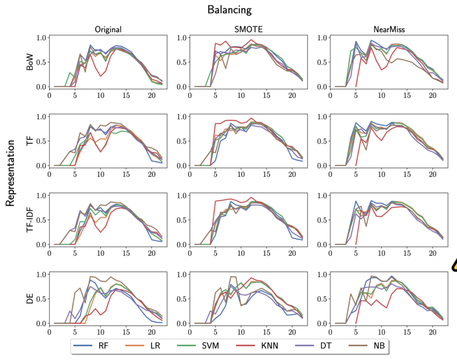
CVE descriptions are enough!
- When CVE descriptions are used, the results are **very similar**.
- Online discussions contribute **marginally**, but they do not harm either.

Emanuele Iannone, Giulia Sellitto, Emanuele Iaccarino, Filomena Ferrucci, Andrea De Lucia, Fabio Palomba

Preprint

Early and Realistic Exploitability Prediction of Just-Disclosed Software Vulnerabilities

RQ2: Impact of Learning Configurations



Zoom: ***CVE + SecurityFocus*** *corpus (F1 score).*

All models share a similar trend
- Very good in central rounds (8th-15th).
- Large drop in final rounds (after 15th).
- **Best round: 15th** (up to 2011).

Best trends (hard to infer)
- Term Frequency
- SMOTE
- Logistic Regression
- **Round 15th**: 0.82 F1 and 0.65 MCC

⚠ In Round 12th, F1 was 0.9+, but MCC was around 0.0!
Reason: The test set had 95% true instances!
Confirmed by the **pessimistic classifier** that had the highest F1.

Emanuele Iannone, Giulia Sellitto, Emanuele Iaccarino, Filomena Ferrucci, Andrea De Lucia, Fabio Palomba

Preprint

Early and Realistic Exploitability Prediction of Just-Disclosed Software Vulnerabilities

The Road Ahead

Is ExploitDB a reliable source for the ground truth?

ExploitDB is very popular, but it cannot keep up with the pace of disclosed vulnerabilities.

Other (**very small**) catalogs: Symantec Attack Signatures, Google Project Zero, Known Exploited Vulnerabilities (KEV), etc.

☛ Aggregating them with ExploitDB would not convert many "Neutral" cases into "Exploitable" (not counting the overlaps with ExploitDB).

Call to action: Enhance the **tracking of exploited vulnerabilities** for a more reliable ground truth.

Future Development
☛ *For better performance*: Measure the **quality of early data** (e.g., readability) and use it to compute the loss more fairly.
☛ *For practical usefulness*: **Combine early** prediction models **with late** models and validate with developers.

Emanuele Iannone, Giulia Sellitto, Emanuele Iaccarino, Filomena Ferrucci, Andrea De Lucia, Fabio Palomba

Preprint

TUHH Hamburg University of Technology

MUR Ministero dell'Università e della Ricerca

SERICS SECURITY AND RIGHTS IN THE CYBERSPACE

Emanuele Iannone

emanuele.iannone@tuhh.de

https://emaiannone.github.io/

@EmanueleIannon3

Early and Realistic Exploitability Prediction of Just-Disclosed Software Vulnerabilities

How Reliable Can It Be?

Preprint