



Improving vulnerability-scores: Leakage-Free Splits and Consistent CVSS 4.0 Scoring

Technical report on the dataset of the VLA severity classifiers

Cédric Bonhomme

cedric.bonhomme@circl.lu

CIRCL - Computer Incident Response Center Luxembourg

2026-10-29

DOI [10.5281/zenodo.23760802](https://doi.org/10.5281/zenodo.23760802)

Contents

1 Summary	2
2 Background	3
3 Finding 1: twin rows across the split	3
3.1 Cause	3
3.2 Measurement on revision 34e6a14	3
3.3 What it means for the previous model	4
4 Finding 2: two readings of CVSS 4.0	6
4.1 Cause	6
4.2 Example	6
4.3 Extent	6
4.4 Choice of reading	6
5 Improvements	7
5.1 Group-aware split	7
5.2 Data cleaning	7
5.3 CVSS 4.0 reading	8
5.4 Validation split	8
5.5 Leakage check	8
5.6 Smaller fixes	8
6 The rebuilt dataset	8
7 Retraining on the new split	9
8 Next steps	10
9 Acknowledgements	11
10 Feedback	11
11 Funding	11
12 How to cite	12
13 References	12

This report describes a new version of the [CIRCL/vulnerability-scores](#) dataset, which trains the VLAI severity classifiers (Bonhomme & Dulaunoy, 2025). The new version gives the models a stricter, more reliable evaluation and more consistent labels. It is the result of a careful external review by Christian Bucher, who measured two problems in the previous version, contributed the patches that fix them and a leakage check that now guards every rebuild. Every figure he reported was reproduced independently before the changes were adopted.

1 Summary

What is new:

- **A leakage-free split.** The dataset is now split by vulnerability group: a CVE record and its GitHub advisory, or two documents with the same text, are always on the same side. The rebuilt dataset has no group shared between train, validation and test, and an automatic check verifies this after every rebuild.
- **Consistent CVSS 4.0 labels.** Every CVSS 4.0 score is now read the same way, as CVSS-B (base metrics only), whatever the source. The same vulnerability always gets the same severity label.
- **A validation split.** Checkpoints are selected on validation, so the test split is used only once, for the final score.
- **Cleaner data and reproducible splits.** Repeated documents and rejected CVEs are removed, and a vulnerability keeps its split from one rebuild to the next.

What we learned:

- In the previous version (revision 34e6a14), 58.45% of the labelled test rows had a train row with the same description, mostly a CVE record and the GitHub advisory for the same CVE (74.6% when shared identifiers and aliases are counted too).
- The previous RoBERTa classifier scores 81.17% on that test split. On the test rows it had never seen in any form, it scores **73.02%**, which is the more representative figure of how it performs on new vulnerabilities.
- CVSS 4.0 scores were read two ways depending on the source, so the same vulnerability could get two different labels. This is now fixed.
- The new RoBERTa classifier, trained on the new dataset and published on the Hugging Face Hub, scores **74.15%** on its leakage-free test split (macro F1 0.646), in line with the estimate above: the new split measures what the model does on new vulnerabilities.

2 Background

`vulntrain-dataset-generation` builds CIRCL/vulnerability-scores from the records held by [Vulnerability-Lookup](#): CVE records (`cvelistv5`), GitHub advisories (`github`), PySec advisories (`pysec`) and CSAF documents from Red Hat, Cisco and CISA. Each row has a description and the CVSS scores found in the record.

`vulntrain-train-severity-classification` maps each row to a severity label from its first available CVSS score, in the order v4.0, v3.1, v3.0, v2.0:

Label	CVSS score
Low	< 4.0
Medium	4.0 – 6.9
High	7.0 – 8.9
Critical	≥ 9.0

Before this change, the dataset was split with a plain random `train_test_split(test_size=0.1)`, with no seed, no grouping and no deduplication.

3 Finding 1: twin rows across the split

3.1 Cause

Vulnerability-Lookup aggregates many sources, so the same vulnerability is often present several times with the same text. A GitHub advisory usually copies the description of its CVE, Red Hat publishes several errata for one CVE, and some CSAF documents were listed twice in the source index. A random row split places such copies on both sides about as often as on the same side, so part of the test score measured how well the model recognises rows it was trained on, rather than how it handles new ones.

3.2 Measurement on revision 34e6a14

Revision 34e6a14 has 740,819 train rows and 82,314 test rows, of which 629,853 and 70,117 have a severity label. Labelled test rows with a labelled train row of the same normalised description (lower case, whitespace collapsed):

Source	Labelled test rows	With a train twin
All	70,117	58.45%
cvelistv5	38,612	50.07%
github	27,501	68.43%
csaf_redhat	2,529	74.65%
csaf_cisa	606	71.62%
pysec	471	65.18%
csaf_cisco	398	51.76%

Counting a shared identifier or alias (for example a GitHub advisory whose aliases list the CVE of a train row) as well, 74.6% of the labelled test rows have a twin in train.

The revision also contains 3,263 repeated documents (2,094 csaf_cisa, 1,169 csaf_cisco), 595 of which have their copies in different splits.

These figures were reported by Christian Bucher and reproduced independently with the same script: the per-row flags match his on all 82,314 test rows.

3.3 What it means for the previous model

The test predictions of [CIRCL/vulnerability-severity-classification-roberta-base](#) (model revision 10d9861) on dataset revision 34e6a14 were split by whether each row has a twin in train. No retraining is needed for this. The model was trained on that dataset revision (2026-10-14), and its model card reports the same 81.17% test accuracy.

Test rows	n	Accuracy	95% CI
All	70,117	81.17%	80.87 – 81.45
With a train twin (description)	40,985	85.83%	85.49 – 86.16
Without a train twin (description)	29,132	74.61%	74.11 – 75.11
With a train twin (description, identifier or alias)	52,342	83.93%	83.62 – 84.24

Test rows	n	Accuracy	95% CI
Without a train twin (description, identifier or alias)	17,775	73.02%	72.36 – 73.67

Rows with a twin are 11.22 points more accurate than rows without one (95% CI 10.61 – 11.82). On the rows it had never seen, the model still classifies nearly three vulnerabilities out of four correctly across four classes.

Per-class recall:

Test rows	Low	Medium	High	Critical
All	0.501	0.865	0.809	0.732
With a train twin (description)	0.527	0.901	0.861	0.801
Without a train twin (description, identifier or alias)	0.474	0.802	0.710	0.626

Low recall is about 0.5 with or without a twin, so it is not affected by the twin rows: it reflects the class itself (rare, and close to Medium) and remains a clear target for future work. High and Critical are the classes where the stricter evaluation makes the biggest difference.

The rows without a twin are not a random sample of the test split: they lean towards CVEs that have no GitHub advisory, i.e. fewer open-source packages. 73.02% is therefore an estimate; the retraining below gives the accuracy on the new split.

For reference, Christian Bucher measured a TF-IDF baseline (the settings of VulnTrain's CERT-FR trainer) at 81.8% on the published split, 74.7% with rows sharing a description kept on one side, and 73.1% with CVE/GHSA aliases grouped too. On unseen rows the previous RoBERTa model is on par with this baseline, which gives a concrete reference point for the next models.

4 Finding 2: two readings of CVSS 4.0

4.1 Cause

CVSS 4.0 has a single score field, filled differently by different producers: with the base score only (CVSS-B), or with the score of the full vector including threat metrics such as Exploit Maturity (CVSS-BT, as NVD does). VulnTrain read CVE records' baseScore as published, but computed the score of GitHub, PySec and CSAF advisories from their vector with `cvss.CVSS4(vector).base_score`, which applies every metric in the vector.

4.2 Example

CVE-2025-6876 and GHSA-pq67-v76v-mvgg describe the same vulnerability with the same description and the same vector:

CVSS:4.0/AV:N/AC:L/AT:N/PR:L/UI:N/VC:L/VI:L/VA:L/SC:N/SI:N/SA:N/E:P

Row	Score used	Label
CVE-2025-6876 (CNA baseScore, CVSS-B)	5.3	Medium
GHSA-pq67-v76v-mvgg (full vector, CVSS-BT)	2.1	Low

The two values match FIRST's CVSS v4.0 calculator.

4.3 Extent

Christian Bucher found that the CVE records are not consistent either: in the 45,432 cvelistV5 records for 2025, at least 586 of the 12,498 CNA CVSS 4.0 blocks publish a baseScore with Exploit Maturity applied, mostly from GitHub and QNAP as CNA. Read as CVSS-B, 421 of them change severity band. 79 of these CVEs are rows of the 34e6a14 test split, 57 of them with a different label under CVSS-B.

4.4 Choice of reading

Every CVSS 4.0 score is now computed from its vector with **base metrics only (CVSS-B)**, for CVE records too. Two reasons:

- Exploit Maturity cannot be inferred from a description, and it changes over time as exploits appear. A model predicting severity from text should predict the intrinsic severity.
- The CVSS 3.x path already ignores temporal metrics.

CVSS4_BASE_ONLY = False in `vulntrain/utls.py` switches every row to the full-vector reading instead. Neither setting reproduces the old mixed reading. CVSS 2.0 and 3.x scores are unchanged.

5 Improvements

All improvements are in [VulnTrain 3.4.0](#).

5.1 Group-aware split

`vulntrain/datasets/grouping.py` puts rows into one group when they share a normalised description, an identifier or an alias, transitively (union-find). Aliases come from the OSV `aliases` field (GitHub, PySec) and from `vulnerabilities[].cve` (CSAF). A record naming more than one CVE (a vendor roll-up, a kernel erratum) does not link through its aliases, so that unrelated vulnerabilities are not chained into one group.

Each group is assigned to train, validation or test from `sha256(seed:smallest_identifier)`. The split therefore does not depend on the order in which records are read, and a vulnerability keeps its split when the dataset is rebuilt, unless its group changes.

New options of `vulntrain-dataset-generation`: `--test-size` (default 0.1), `--validation-size` (default 0, no validation split) and `--seed` (default 42). Rows now have an `aliases` column.

On revision 34e6a14 (descriptions and identifiers only, no alias data), the grouping gives 467,487 groups; the largest has 881 rows, and the test share is 9.98%, so grouping does not unbalance the split.

5.2 Data cleaning

- Repeated documents (the same source and identifier) are kept once.
- Rejected CVE records are skipped. They only carry a boilerplate “Rejected reason” description (3,197 rows in 34e6a14, 9 of them labelled) and formed the largest description groups.

5.3 CVSS 4.0 reading

See [Choice of reading](#).

5.4 Validation split

When the dataset has a `validation` split, the severity trainer selects the best checkpoint on it and evaluates the test split once at the end, so the test score is not used for model selection.

5.5 Leakage check

`tools/dataset/check_split_leakage.py` (contributed by Christian Bucher) reports, for each split other than `train`, the labelled rows that have a train twin by description or by identifier/alias, and the groups that span more than one split. With `--max-shared-groups 0` it exits with status 1 when any group spans two splits, so it can be run after each rebuild. See `tools/dataset/README.md`.

5.6 Smaller fixes

- `compute_metrics` no longer fails when an evaluation split and its predictions both lack a severity class.
- Every trainer accepts `--no-push`, and `--repo-id` is then optional.

6 The rebuilt dataset

Published as revision [55a886a](#) of `CIRCL/vulnerability-scores` on 2026-10-29, built with `--validation-size 0.1` (80/10/10). It is identical to the build the model of the next section was trained on.

Split	Rows	Labelled rows
train	670,283	567,964
validation	83,550	71,470
test	83,376	70,472
Total	837,209	709,906

The total is higher than in 34e6a14 (823,133 rows): new records since then, minus the repeated documents and the rejected CVEs.

Leakage check on the rebuilt dataset: no labelled test or validation row has a train twin by description, identifier or alias; no description group and no vulnerability group spans two splits; no identifier is repeated within a source.

Labelled test rows by source (share of the test split in brackets):

Source	Labelled test rows
cvelistv5	38,817 (55.1%)
github	27,814 (39.5%)
csaf_redhat	2,717 (3.9%)
pysec	509 (0.7%)
csaf_cisa	343 (0.5%)
csaf_cisco	272 (0.4%)

The source mix of the test split is close to that of 34e6a14.

7 Retraining on the new split

The new [CIRCL/vulnerability-severity-classification-roberta-base](#) (model revision 064daa2) was trained on dataset revision 55a886a, with roberta-base and the same hyperparameters as the previous model (learning rate 3e-5, per-device batch size 8, 5 epochs, weight decay 0.01, best checkpoint by validation accuracy), on 4 × NVIDIA L40S. This training configuration and hardware are described in detail in Bonhomme (2025).

The best checkpoint on validation is the one of epoch 4 (validation accuracy 75.00%). It was evaluated once on the 70,472 labelled rows of the test split. Training took 4 h 25 min and used 4.61 kWh (3.99 kWh for the GPUs), for 0.48 kg CO₂eq, as measured by CodeCarbon on the Luxembourg grid.

Model	Evaluated on	Accuracy	95% CI	Macro F1
Previous RoBERTa (10d9861)	34e6a14 test split (leaky)	81.17%	80.87 – 81.45	–

Model	Evaluated on	Accuracy	95% CI	Macro F1
Previous RoBERTa (10d9861)	34e6a14 test rows without a twin	73.02%	72.36 – 73.67	–
New RoBERTa (064daa2)	55a886a test split (leakage-free)	74.15%	73.83 – 74.47	0.646

The new model scores about one point above the previous model on the rows it had never seen. The two test sets are not identical (the labels now use CVSS-B and rejected CVEs are gone), but the agreement confirms that the leakage-free split gives the honest figure directly, without the need for a twin analysis: about three vulnerabilities out of four are classified correctly across four classes. An earlier training run on the same data scored 73.79% (macro F1 0.650), which gives an idea of the run-to-run variation.

Per-class results of the new model:

Class	Precision	Recall	F1
Low	0.615	0.328	0.428
Medium	0.777	0.825	0.800
High	0.733	0.724	0.728
Critical	0.646	0.611	0.628

Medium, High and Critical recall are close to those of the previous model on the rows without a twin (0.802, 0.710 and 0.626). Low recall is lower (0.328 against 0.474): Low is the rarest class, and it is now the main target for the next models.

Validation accuracy levels off after the second epoch (72.3%, 74.4%, 74.6%, 75.0%, 74.6%) while validation loss rises from 2.54 to 3.01, so a shorter training would likely give the same accuracy for less compute.

8 Next steps

- The other severity classifiers are retrained on the new dataset. Their model cards will report accuracy on the leakage-free split, together with the dataset revision they were trained on.

Figures reported on the previous split are about 7 points higher for RoBERTa than the leakage-free figures, and are superseded by the new ones.

- The leakage check runs after every rebuild of the dataset, so future versions keep the same guarantee.
- Earlier revisions of the dataset stay available on the Hugging Face Hub, so earlier results remain reproducible by citing their revision.
- Low-class recall (0.328 on the new split), the comparison with a TF-IDF baseline and a shorter training schedule give clear targets for the next models.

9 Acknowledgements

Many thanks to Christian Bucher for his careful and constructive review: the analysis, the patches (group-aware split, CVSS 4.0 reading, validation split), the per-row twin flags, and the leakage and seen/unseen evaluation scripts.

10 Feedback

Feel free to share your feedback at info@circl.lu or publicly: <https://github.com/vulnerability-lookup/VulnTrain/issues>

11 Funding



**Co-funded by
the European Union**

AIPITCH (AI-Powered Innovative Toolkit for Cybersecurity Hubs) is a co-funded EU project supported by the European Cybersecurity Competence Centre (ECCC) under the DIGITAL-ECCC-2024-DEPLOY-CYBER-06-ENABLINGTECH program and [CIRCL](#).

Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Cybersecurity Competence Centre. Neither the European Union nor the European Cybersecurity Competence Centre can be held responsible for them.

12 How to cite

Bonhomme, C. (2026). *Improving vulnerability-scores: Leakage-Free Splits and Consistent CVSS 4.0 Scoring*. Technical report, CIRCL. <https://doi.org/10.5281/zenodo.23760802>

```
@techreport{bonhomme2026vulnscores,  
  title      = {Improving vulnerability-scores: Leakage-Free Splits and  
                Consistent {CVSS} 4.0 Scoring},  
  author     = {Bonhomme, C{\`e}dric},  
  institution = {CIRCL - Computer Incident Response Center Luxembourg},  
  year       = {2026},  
  month      = oct,  
  doi        = {10.5281/zenodo.23760802},  
  url        = {https://doi.org/10.5281/zenodo.23760802}  
}
```

13 References

- Bonhomme, C., & Dulaunoy, A. (2025). *VLA: A RoBERTa-Based Model for Automated Vulnerability Severity Classification*. arXiv preprint arXiv:2507.03607. <https://arxiv.org/abs/2507.03607>
- Bonhomme, C. (2025). *GPU Efficiency in VLA Model Training: Experiences and Benchmarks from Months of VLA Vulnerability Severity Classification Model Training*. Technical report, CIRCL. [doi:10.13140/RG.2.2.24233.04963](https://doi.org/10.13140/RG.2.2.24233.04963)