Qeios

Peer Review

Review of: "R+R: Security Vulnerability Dataset Quality Is Critical"

Haonan Li[1]

1. University of California, Riverside, United States

This paper investigates the data quality of the widely-used **VulRepair** security-vulnerability repair corpus and several derivative datasets. The authors demonstrate that (i) massive intra- and inter-split duplication, (ii) mislabeled CWE tags, and (iii) truncated or otherwise incomplete code snippets have led to serious **data-leakage** and **label-noise** problems. Re-evaluating four popular code-repair models (VulRepair, CodeBERT, GraphCodeBERT, UniXcoder) under deduplicated and partially relabeled conditions, they show a significant **accuracy drop**.

Strengths

- **Empirical analysis & Actionable insight:** The authors show the importance of data duplication and provide actionable methods to solve it.
- **Real-world Impact:** This work raises a red flag for dozens of recent relevant papers that may need to revisit their claims; future work is required to consider the issue of data duplication **more carefully**.

Weaknesses/Limitation

- **Not considering GPT-era works:** I remember that recent works, based on GPT-4 or o-series models, reach much better results. Yet this paper doesn't discuss these results.
- **Shallow duplicate detection:** The authors rely on exact or near-exact string hashing. But modern LLMs may leverage *semantic* clones—code that differs lexically yet implements the same pattern—so more sophisticated clone-detection is required to get a much cleaner view.

In summary, this work convincingly shows that a large portion of the reported progress in LLM-based vulnerability repair is an illusion created by dataset leakage and noisy labels. While some methodological

 doi.org/10.32388/DV2EKQ

corners remain to be tightened, the authors' dataset audit and re-evaluation provide an essential service to the community.

Declarations

Potential competing interests: No potential competing interests to declare.