Measuring What Matters: Systematic Code Quality Assessment for Software Vulnerability Datasets

Arthur Dunbar, Rakesh Verma, Amutheezan Sivagnanam
Department of Computer Science, University of Houston

1. The Grand Challenge

- SVD is foundational to software security, directly informing vulnerability triage, patch prioritization, and secure development workflows.
- The reliability and generalizability of ML-based SVD systems hinge critically on the quality of their training and evaluation datasets.
- Quality of dataset contributes towards effective learning of automated vulnerability detection using Machine Learning
- Widely used datasets suffer from duplication, information leakage, and spurious correlations

2. Data Quality

- We introduce a comprehensive taxonomy (See Figure 3) and introduce corresponding data quality metrics to access the software vulnerability datasets (SVD).
- We perform evaluation across the data quality metrics using all 12 datasets **(i.e. BigVul (BV), Devign (Dev), DiverseVul (Div), D2A, MegaVul (MV), MVDSC (MVD), ReVeal (ReV), RealVul (Real), and VulDeePecker (VDP), PrimeVul (PV), SecEvalVul (Sec), TitalVul (TV)).**

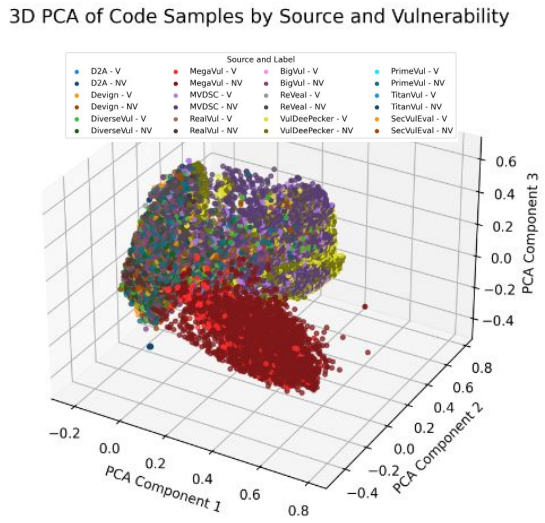

Figure 1. 3D PCA visualization of TF-IDF-encoded functions from the chosen 12 datasets. Color–coded by vulnerability label, the plot shows clustering patterns that indicate separability of vulnerability and non-vulnerability functions.

Figure 2. Relationship between foundational data and derived datasets. This visualization highlights how code projects like QEMU and FFMpeg (shown on the right) serve as the foundational sources for widely used datasets like DiverseVul and MegaVul (shown on the bottom). Each row is a repository that is one of the top 20 most common repositories in a dataset. Each cell is filled in if that dataset (bottom) has that repository (right). The blue colors represent how common the repo is in the dataset with darker shades indicating more common occurrences. Gray cells indicate a lack of data for determining prevalence.

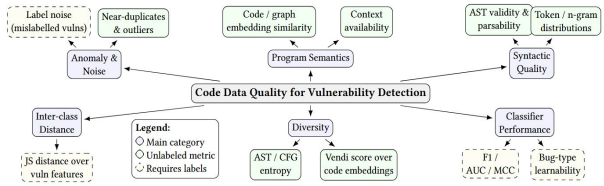

Figure 3. Data-quality taxonomy for code datasets targeting software vulnerability detection.

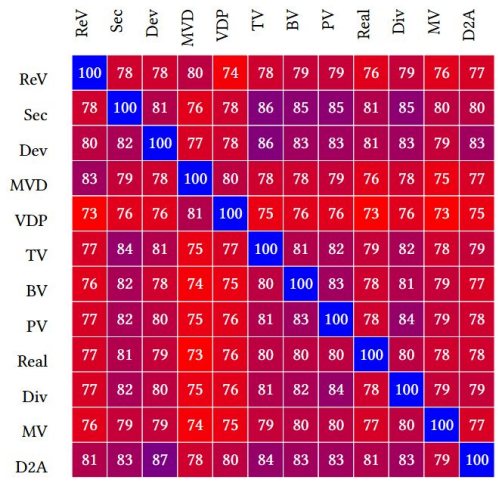

Figure 4. Cosine Similarity comparisons of code samples using embeddings generated by CodeT5+-16B from the 12 datasets considered, values are multiplied by 100 for better readability. 0.73 is the bottom of the scale. As a baseline for similarity comparisons, 10,000 random instances were taken from each dataset and paired against examples from the other datasets. This resulted in an average cosine similarity of 0.65.

Dataset Model	ReV	Sec	Dev	MVD	VDP	TV	BV	PV	Real	Div	MV	D2A*	Avg
LR*	11.57	16.68	53.80	81.72	73.76	28.07	11.82	1.26	19.40	4.16	0.48	4.09	25.57
Ada*	31.46	11.24	41.93	74.21	74.21	45.56	11.71	4.08	36.24	5.60	3.96	2.57	28.56
RF*	13.98	16.59	44.88	85.23	85.35	29.57	25.33	11.84	48.24	14.84	8.52	19.99	33.70
SVM	14.18	9.56	52.75	89.08	86.47	21.83	21.87	3.47	42.25	8.41	36.29	13.32	33.29
DB*	35.22	37.01	57.65	94.75	95.32	47.97	36.32	10.7	99.58	20.29	46.78	56.57	53.18
CB*	41.49	33.85	57.71	94.40	95.94	49.62	83.48	20.48	99.76	22.68	47.05	57.10	58.63
GCB*	41.38	35.04	57.92	94.16	96.03	49.05	84.14	23.57	99.78	22.81	48.70	57.97	59.21
CT5+*	30.48	36.70	58.71	89.99	78.08	53.76	86.46	39.81	99.90	30.51	43.76	50.44	58.22
Avg	27.47	24.58	53.17	87.94	85.65	40.68	45.14	14.40	68.14	16.16	29.44	32.76	43.79
Score	0.77	0.75	0.97	0.62	0.64	0.91	0.95	0.64	0.82	0.66	0.79	0.83	-

Table 1. F1 Scores for Classical and Deep Learning Models across datasets. *LR-Logistic Regression, Ada-Adaboost, RF-Random Forest, DB-DistilBERT, CB-CodeBERT, GCB-GraphCodeBERT, CT5+-CodeT5+-16B
*D2A used a stratified sample of 100,000 instances

Dataset	CC	Class	G	DC	CCmp	JSD	ConC	VS	ICD	Ren	Par
ReV	0.41	0.77	0.25	0.93	0.94	0.81	0.62	0.55	0.71	0.72	0.68
Sec	0.91	0.75	**0.66**	0.69	0.90	0.06	0.70	0.29	0.48	**0.73**	0.77
Dev	0.90	**0.97**	0.29	0.54	0.97	0.67	**0.85**	0.56	0.48	0.72	0.80
MVD	0.91	0.62	0.19	**0.98**	**1.00**	0.19	0.39	0.04	0.50	0.68	0.17
VDP	0.84	0.64	0.23	**0.98**	0.46	0.51	0.45	0.05	-	-	-
TV	0.96	0.91	0.49	0.67	0.84	0.01	0.71	0.29	0.34	0.72	0.76
BV	0.31	0.95	0.18	**0.98**	0.91	0.51	0.70	0.90	0.49	0.72	**0.86**
PV	0.22	0.64	0.42	0.97	0.86	**0.99**	0.76	**0.98**	**0.86**	0.71	0.80
Real	0.16	0.82	0.06	**0.98**	0.72	0.00	0.81	0.42	0.50	0.67	0.29
Div	0.33	0.66	0.45	0.94	0.84	0.92	0.77	**0.98**	0.57	0.70	0.78
MV	0.23	0.79	0.14	0.95	0.95	0.60	0.75	0.38	0.80	0.71	0.78
D2A	**0.98**	0.83	0.07	0.86	0.76	0.70	0.55	0.17	0.75	0.68	0.83

Table 2. Overall Dataset Score Results. CC - Char Count, Class - Classification, G - Generalization, DC - Data Cartography, CCmp - Consistency and Completeness, ConC - Construct Coverage, VS - Vendi Score, ICD - Inter Class Distance, Ren - Renyi Entropy, Par - Parsability. Bolded Values are the highest in their column.

4. Results & Insights

- Most datasets are either too noisy, and rarely generalize across one another.
- Automated label-quality tools detect near-duplicates reliably but largely fail to catch mislabeled samples.
- D2A contains cross-class duplicates, RealVul scores poorly on completeness, and multi-block samples break AST parsers in RealVul and MVDSC.
- Distributional analyses flag TitanVul, SecVulEval, and RealVul as outliers, while PrimeVul and MegaVul show the strongest vulnerable/non-vulnerable class separation.