1. On the Impact of Sample Duplication in Machine-Learning-Based Android Malware Detection (*Open Access*)

Zhao, Yanjie (1); Li, Li (1); Wang, Haoyu (2); Cai, Haipeng (3); Bissyandé, Tegawendé F. (4); Klein, Jacques (4); Grundy, John (1)

Source: *ACM Transactions on Software Engineering and Methodology*, v 30, n 3, May 2021, *10.1145/3428015*; **ISSN:** 1049331X, **E-ISSN:** 15577392; **DOI:** 10.1145/3446905; **Article number:** 40; **Publisher:** Association for Computing Machinery

Author affiliation: (1) Monash University, Clayton, Australia (2) Beijing University of Posts and Telecommunications, Beijing, China (3) Washington State University, United States (4) University of Luxembourg, Luxembourg

Abstract: Malware detection at scale in the Android realm is often carried out using machine learning techniques. State-of-the-art approaches such as DREBIN and MaMaDroid are reported to yield high detection rates when assessed against well-known datasets. Unfortunately, such datasets may include a large portion of duplicated samples, which may bias recorded experimental results and insights. In this article, we perform extensive experiments to measure the performance gap that occurs when datasets are de-duplicated. Our experimental results reveal that duplication in published datasets has a limited impact on supervised malware classification models. This observation contrasts with the finding of Allamanis on the general case of machine learning bias for big code. Our experiments, however, show that sample duplication more substantially affects unsupervised learning models (e.g., malware family clustering). Nevertheless, we argue that our fellow researchers and practitioners should always take sample duplication into consideration when performing machine-learning-based (via either supervised or unsupervised learning) Android malware detections, no matter how significant the impact might be. © 2021 ACM. (77 refs)

Main heading: Mobile security
Controlled terms: Android (operating system) - Classification (of information) - Large dataset - Learning systems - Malware - Unsupervised learning
Uncontrolled terms: Android malware - High detection rate - Machine learning techniques - Malware classifications - Malware detection - Malware families - Performance gaps - State-of-the-art approach
Classification Code: 716.1 Information Theory and Signal Processing - 723 Computer Software, Data Handling and Applications - 723.2 Data Processing and Image Processing
Funding Details: Number: 830892, Acronym: H2020, Sponsor: Horizon 2020 Framework Programme; Number: DE200100016,DP200100020,FL190100035, Acronym: ARC, Sponsor: Australian Research Council; Number: 61702045,62072046, Acronym: NSFC, Sponsor: National Natural Science Foundation of China; Number: CHARACTERIZE C17/IS/11693861, Acronym: FNR, Sponsor: Fonds National de la Recherche Luxembourg;
Funding text: This work was supported by the Australian Research Council (ARC) under a Laureate Fellowship project FL190100035; a Discovery Early Career Researcher Award (DECRA) project DE200100016; and a Discovery project DP200100020; by the National Natural Science Foundation of China (No. 61702045 and No. 62072046); by the Fonds National de la Recherche (FNR), Luxembourg, under project CHARACTERIZE C17/IS/11693861; and by the SPARTA project, which has received funding from the European Union's Horizon 2020 research and innovation program under grant agreement No. 830892. Authors' addresses: Y. Zhao, L. Li (corresponding author), and J. Grundy, Monash University, Australia, Wellington Rd, Clay-ton, VIC, 3800; emails: {yanjie.zhao, li.li, john.grundy}@monash.edu; H. Wang, Beijing University of Posts and Telecommunications, China, No 10, Xitucheng Road, Haidian District, Beijing, PRC, 100876; email: haoyuwang@bupt.edu.cn; H. Cai, Washington State University, USA, Pullman, WA 99163, USA; email: haipeng.cai@wsu.edu; T. F. Bissyandé and J. Klein, University of Luxembourg, Luxembourg, 2 Avenue de l'Universite, 4365 Esch-sur-Alzette, Luxembourg; emails: {tegawende.bissyande, jacques.klein}@uni.lu. 1049-331X/2021/05-ART40 $15.00 https://doi.org/10.1145/3446905
Open Access type(s): All Open Access, Green
Database: Compendex
Data Provider: Engineering Village