Analysing mobile forensic datasets: A systematic review on availability, efficacy, and limitations

Monika[1], *Yogesh* K. Sharma[1], *Deepak* S. Tomar[1], and *R. K.* Pateriya[1]

[1]Maulana Azad National Institute of Engineering and Technology, Department of Computer Science and Engineering, Bhopal, M.P, India

Abstract. Everyday there is an increase in the number of malwares being created which presents a significant danger to the Android systems holding a large share in the operating systems market. This surge in malware creation also makes it challenging to analyse and detect these malicious applications. Machine learning techniques are commonly used for malware detection, but the development of an effective system requires a reliable dataset to train and test it. This paper provides an overview of the most commonly used datasets in malware detection research conducted between 2015-2020, based on their performance, usability, availability, and effectiveness. By analysing and comparing these datasets, this paper aims to provide insights into the selection of appropriate datasets for future research in this area.

1 Introduction

With the proliferation of mobile devices and applications, the number of mobile malwares has been increasing rapidly. Numerous research initiatives have started to address this issue as a result of the rise in mobile malware, which has raised concerns about the security of mobile apps. To stop the spread of harmful software on mobile devices, researchers have been attempting to create efficient malware detection tools. These include approaches based on static analysis, dynamic analysis and approaches based on hybrid analysis, which combine the advantages of both static and dynamic analysis techniques.

In static analysis, features are retrieved from Android applications without executing them on a real device or Android emulator [1]. Static analysis has several benefits, including code coverage and rapid feature extraction, while in dynamic analysis the functionalities are retrieved by executing the applications on an Android emulator or an actual Android device [2]. Applications are run in a controlled environment like an android emulator which simulates android devices and is mainly used for running and testing the applications. Some other advantages of static analysis include quick processing and reduced resource usage, such as memory and CPU [3]. These benefits result from the fact that this technique reverse engineers the applications to thoroughly examine the code without running the program.

It is critical to have access to a reliable dataset of malware samples that is current and indicative of the current landscape of Android malware in order to build a strong malware detection system and properly measure its performance. The collection needs to include in-depth details on the behaviour and strategies malware developers use to carry out their

nefarious deeds. The dataset must have extensive behaviour profiles for each form of malware in order to fully comprehend the harmful strategies employed in modern malware apps.

Researchers need either access to the technique/tool or the data source itself in order to compare or improve results using the same input data sources. Researchers often prefer access to datasets since testing may be time-consuming and require expertise with an unfamiliar methodology. Therefore, having readily available datasets is essential to ensuring accessibility.

Various Android malware datasets covering the period from 2010 to 2020 have been assessed in this paper to determine their availability and analyse their shortcomings and limitations. Furthermore, this article presents a summary of the dataset's performance and popularity used in research carried out from 2015 to 2020. By addressing the limitations of existing benchmarks, researchers can develop more robust and effective malware detection systems to combat the evolving landscape of Android malware.

2 Data collection and data generation

An Android malware dataset consists of Android applications collected from various sources such as the Google Play Store, online repositories, packages, and already existing datasets. These applications are flagged as either malicious or benign, which can be done with the help of online malware detection engines, software, publicly released reports by security companies, and other methods.
In this section, some of the approaches to create a dataset have been discussed.

A well-known and extremely trustworthy dataset for identifying Android malware is Rmvdroid [4]. It was made by utilizing a crawler to gather metadata from Google Play's app maintenance results over a period of many years. In order to detect malware, the dataset also used Google Play's app maintenance procedures. The first four screenshots of Google Play, in particular, were captured in 2014, 2015, 2017, and 2018, respectively. Applications reported by VirusTotal as exhibiting potentially sensitive behaviours were followed closely on Google Play to see whether any action had been taken against them by Google. Any sensitive or suspicious apps that Google Play deleted throughout the monitoring period were marked as dangerous apps. Through this methodology, a highly extensive dataset comprising of 9,133 samples of malware, linked to 56 malware families, was generated.

Wei et al. [5] compiled a dataset by collecting sample application programs from various sources, including VirusShare, Google Play, and other security firms, that were developed between 2010 and 2016. Most of the malware in this collection did not have an assigned malware family name, and therefore, the initial step was to identify the family name for such applications. Each application was evaluated by over 55 VirusTotal antivirus programs, with each result being a candidate label or not-a-malware. If at least 50% of the antivirus programs used in VirusTotal identified the application as malware, it was classified as such. Following this method, 1,216,885 of the gathered applications were not classified as malware by any antivirus software, while 195,185 had been classified as malware by some antivirus program but had not yet met the 50% threshold. A total of 52,520 applications were identified as malicious and a malware family was assigned using "dominant keyword algorithm".

3 Methodology

An overview of popular datasets used in the field of Android malware detection created during the period 2010-2020 is provided in this paper. To summarize the popular datasets, research papers published in the period 2010-2020 were taken. The evaluation included

assessing the availability, year of creation, frequency of use, and the source from which these datasets can be downloaded. The analysis revealed that most datasets are publicly available, while others are only accessible upon request to prevent dataset misuse. Some authors have also discontinued the distribution of their datasets due to maintenance issues.

To determine the effectiveness of these datasets, research papers published between 2015 and 2020 were reviewed and the accuracy achieved by the detection systems trained and tested on these datasets was analysed.

4 Available datasets for malware detection and their limitations

The availability of data is crucial for research as it facilitates reproducible results and enables the improvement of the state of the art. Consistent input data sources are necessary for the comparison and improvement of results. The availability of several datasets for the purpose of android malware detection have been summarized below in Table 1.

Table 1. Summary of availability of datasets used in the study

Dataset availability	Dataset	Percentage
Available publicly	Drebin [6], AndroTracker [7], SAPPIMMDS [7], Andro-Profiler [7], Andro-Dumpsys [7], Kharon[8], AAGM [9], Kronodroid [10], CICAndroidBot [11], CICAndMal2017 [12], CICAndMal2020 [13], Contagio dump [14]	66.67%
Available on request	RmvDroid [4], AndroZoo [15]	11.11%
Not available	AMD [5], The Genome Project [16], PRAGuard dataset [17]	22.23%

4.1 The Drebin dataset

The Drebin dataset is a publicly accessible resource and is widely acknowledged as a highly referenced dataset in the research and studies related to Android malware. Drebin dataset consists of 5560 malware samples across 179 different malware families and 123,453 benign samples collected from 2010 to 2012 [6]. The samples were gathered over a period of 2 years from August 2010 and October 2012 and can be accessed via the Mobile-Sandbox Project. Drebin dataset provides various features including API calls, network addresses, hardware features, intents, permissions, and various components of the application.

Nearly half (49.35%) of the dataset contains repackaged Android apps with identical opcode sequences [18], created by disassembling and adding malicious payloads before repackaging and distributing them on Android markets. As a result, the test set includes many duplicates from the training set, potentially inflating algorithm performance and leading to erroneous research conclusions.

4.2 AndroZoo dataset

AndroZoo [15] is an ever-expanding archive of Android applications accumulated from multiple sources, such as the official Google Play app store. Initially created in 2016, it contained more than three million applications. At present, the dataset comprises more than 21 million unique APKs, each of which has either undergone or will undergo scrutiny by several antivirus engines to ascertain which ones are classified as malicious software. To detect potential malware, the applications included in the AndroZoo dataset are analysed and

categorized by employing more than 60 antivirus tools [19]. However, the AndroZoo dataset lacks explicit static features, requiring researchers to extract the features themselves.

AndroZoo solely offers the count of antivirus flags received by an application on the online portal for VirusTotal. While the main AndroZoo file notes the number of flags, it doesn't offer details on the flagging antivirus software.

4.3 The Genome Project

Collected for a span of a little over a year, between August 2010 and October 2011, this dataset encompasses 1,260 instances of Android malware in 49 distinct malware families, providing coverage for the majority of present Android malware families. The items in the dataset were collected using manual and automated crawling from various Android markets. It contains static features such as permissions, intents, etc. The sharing and download of this dataset came to a halt in the year 2015 due to limited resources and graduation of students who were involved with the Android Malware Genome Project [16].

4.4 The AMD dataset

The AMD dataset [5], which has been accessible since 2017, encompasses 24,553 Android malware applications divided into 71 families. AMD is among the most comprehensive publicly available Android malware datasets, featuring malware instances gathered from 2010 to 2016. A total of 55 antivirus programs that are listed on VirusTotal were used to thoroughly scan the Android Malware Dataset. The VirusTotal tool marks a sample as malware if more than 28 anti-virus programs classify it as such. The methodical approach of the AMD dataset includes assessing malware behaviour and analysing harmful components according to their priorities. Sequences of Dalvik bytecode opcodes, Java VM-type signatures, the values of constant-string instructions, and signatures are a few examples of the parts that make up the characteristics that are contained. Rafiq et al. [19], investigated AMD dataset and found that 29.8% applications in the AMD are repacked malware.

4.5 CICAndMal2017

CICAndMal2017 dataset is made by executing both malicious and benign applications on physical mobile devices to prevent alteration of runtime behaviour by advanced malware samples that can detect the emulator environment. Multiple sources are utilized to gather 4,354 malware samples and 6,500 benign samples, amounting to over 10,854 samples in total. The CICAndMal2017 dataset [12] comprises benign applications obtained from the Google Play store in 2015, 2016, and 2017, as well as malware samples classified into four distinct categories-SMS malware, scareware, adware, and ransomware.

4.6 The Kharon dataset

The dataset contains samples from well-known Android Malware sets, including The Genome Project [16] and the Contagio Mobile Dataset [14]. The Kharon Dataset [8], created by executing seven malware samples on a mobile device, records various features and behaviours of the malware. The AndroBlare tool monitors information flow between system objects and generates a human-readable directed graph for each malware sample, making it easier for researchers to identify patterns and understand their behaviour. The Kharon Dataset is valuable for research on Android malware detection and prevention as it provides a unique perspective on the actions taken by malware on a real mobile device.

5 Analysis on dataset popularity and usability

In this section, the popularity, usability, and performance of various datasets have been analysed. The datasets collected for analysis have been used in several Android Malware detection models from 2015-2020. The result and accuracy of these models have been taken as a measure to determine the performance of datasets.

Table 3. Summary of usage and accuracy of multiple datasets during the period 2015-2020.

Dataset / Data source	Launch Year	Total number of applications	Total number of malicious applications	Average accuracy (%)	Percentage of occurrence (%)
AMD [5]	2017	24,553	24,553	95.33	11.54
Drebin [6]	2014	129,013	5,560	94.12	48.08
AndroZoo [15]	2016	21,000,000+	-	89.1	3.85
Contagio [14]	2018	28,760	11,960	94.45	11.54
Google Play [20]	-	Android app market	-	93.43	28.85
MalGenome [16]	2012	1,260	1,260	95.15	21.15
APKPure [21]	-	Android app market	-	94.77	7.69
VirusShare [22]	2012	Invitation only access	-	97.68	7.69
McAfee [23]	2013	50,926	-	98.3	5.77
Comodo cloud sec. [24]	-	Comodo Sec. Lab Sample	-	95.66	5.77

Among the research papers considered in this study during the period of 2015-2020, Table 2 lists the 11 most frequently used datasets.

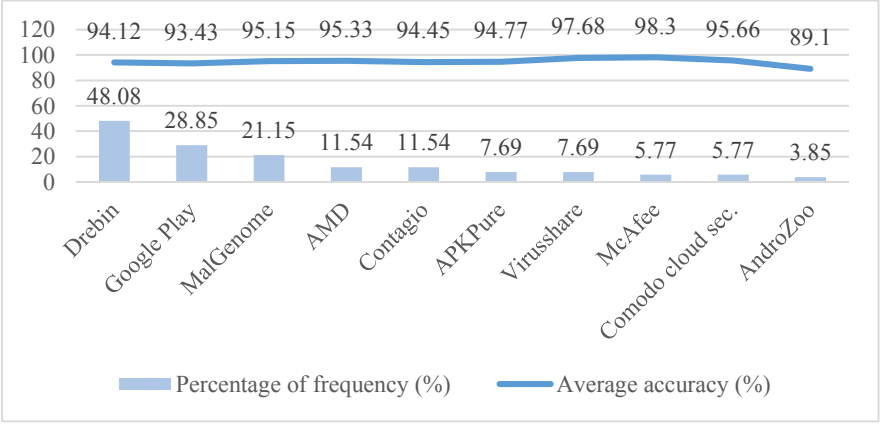

Fig. 1. Dataset popularity and efficiency based on usage frequency in research carried out from 2015-2020, expressed as average accuracy and percentage of overall occurrence.

The Drebin dataset stood out as the most frequently utilized among the analysed datasets, with a usage rate of 48.08% and an average accuracy of 94.12%. Following closely was the dataset of applications collected directly from Google Play, which had an average accuracy of 93.43%, as demonstrated in Fig 1.

6 Data incompleteness in research datasets

The detection of Android malware has been the subject of numerous research studies recently. Therefore, it is crucial to have a dependable and extensive malware dataset to create effective malware classifiers and assess the performance of different detection models. However, even though existing benchmarks for Android malware have been widely used in the research community, they have notable drawbacks that need to be considered.

For example, many applications in the databases are simply copies of malware that has already been found. Malware developers typically repackage already-existing malware applications to produce malware clones with the least amount of effort and cost instead of developing fresh versions [19]. An analysis of diverse datasets by Rafiq et al. [19] unveiled that repackaged malware constitutes 29.8% in AMD, 52.3% of applications in Drebin, and 42.3% in the AndroZoo dataset. Duplicate entries need to be removed from the dataset because they increase overhead in terms of time and computational expenses and can even lead to incorrect results by machine learning models.

Moreover, the majority of the available datasets are out of date and unable to capture recent trends in malware evolution [10]. The influence of time on the performance of machine learning-based classifiers employed in Android malware research has not been well taken into account. This can result in concept drift, in which virus properties change over time and undermine the accuracy of previously reliable classifiers [4].

The malware research community often employs VirusTotal to categorize Android apps based on results from roughly 60 antivirus scanners. However, there are no fixed rules for interpreting scan results, leading to the use of multiple threshold-based labelling methodologies. Typically, apps are labelled harmful if a certain number of scanners identify them as such (e.g., 10 or more). Yet, there is no standard for determining the appropriate cut off for labelling an app as malicious. While some threshold values can approximate actual malware, VirusTotal's frequent rotation of scanners makes such thresholds inaccurate in the long term. As Android malware grows more complex and evolves, relying solely on VirusTotal scans to label an app as malicious may not fully capture the true nature of Android malware.

Arp et al. [6] used a threshold of 2 antivirus flags to label an app as malicious. However, Li et al. [25] used a threshold of 28, and Rafiq et al. [19] used the criteria of 10 antivirus flags. Comprehensive manual research is necessary to gather precise and complete information on malware behaviours, despite the usefulness of sources such as VirusTotal.

A major issue in Android malware research is the lack of dataset availability, which impedes replication and scientific advancement. Sharing data is crucial in science, but some datasets, such as the AMD dataset and Android Malware Genome, are no longer accessible due to resource constraints. This lack of availability restricts researchers' ability to validate and expand upon prior studies, hindering scientific progress. To enhance reproducibility and promote science, it is essential to ensure that datasets are shared and accessible to researchers.

7 Discussion and conclusion

The creation of high-quality datasets involves various considerations regarding the definition of a "perfect" dataset and the standards for identifying benign and malicious applications. A review of existing datasets has revealed that limitations and shortages can be present in any given dataset, which presents a major obstacle for advancing research and development in detecting Android malware.

Unavailability of datasets can hinder the replication of results and the improvement of current methodologies. Approximately 22.23% of all datasets analysed are not publicly available, while 11.11% are only accessible upon request to prevent dataset misuse. Some

authors have also discontinued the distribution of their datasets due to maintenance issues. Another issue is application repackaging which results in copies of applications that are already present in the database. This issue can lead to incorrect results and increase the computational overhead. Lastly, dataset ageing is a significant setback for research in forensics, as with time malware is evolving, making existing datasets less relevant and less accurate. Addressing these deficiencies is essential for the creation of effective and reliable malware detection models.

Among all the datasets analysed, the popularity of the Drebin dataset (appearing 48% of the times with 94.12% accuracy) demonstrates its usefulness and efficacy for researchers in detecting Android malware. Further efforts should be made to address the gaps in existing datasets and develop new datasets that accurately reflect the evolving landscape of Android malware.

References

1. Pan, Ya, X. Ge, C. Fang, Y. Fan. *A systematic literature review of android malware detection using static analysis.* IEEE Access 8, 116363-116379 (2020).
2. Amin, M. Rakib, M. Zaman, M. S. Hossain, M. Atiquzzaman. *Behavioral malware detection approaches for Android.* In 2016 IEEE International Conference on Communications (ICC), IEEE, pp. 1-6 (2016).
3. Chess, Brian, and G. McGraw. *Static analysis for security*. IEEE Sec. & Prvcy **2**, 76 (2004).
4. Wang, Haoyu, J. Si, H. Li, Y. Guo, *Rmvdroid: towards a reliable android malware dataset with app metadata*, In IEEE/ACM 16th International Conference on Mining Software Repositories (MSR), 404 (2019).
5. Wei, Fengguo, Y. Li, S. Roy, X. Ou, W. Zhou, *Deep ground truth analysis of current android malware*, in International Conference on Detection of Intrusions and Malware, and Vulnerability Assessment, Springer, Cham, 252 (2017).
6. Arp, Daniel, M. Spreitzenbarth, M. Hubner, H. Gascon, K. Rieck, C. E. R. T., *Drebin: Effective and explainable detection of android malware in your pocket*, Ndss, 23 (2014).
7. *HCRL, Datasets for malware/malicious app analysis*. https://ocslab.hksecurity.net/Datasets (accessed February, 05 2023).
8. Kiss, Nicolas, J. F. Lalande, M. Leslous, V. V. T. Triem Tong, *Kharon dataset: Android malware under a microscope*, in The LASER Workshop: Learning from Authoritative Security Experiment Results, LASER, 1 (2016).
9. A. H. Lashkari, A. Fitriah A. Kadir, H. Gonzalez, K. F. Mbah, A. A. Ghorbani, *Towards a Network-Based Framework for Android Malware Detection and Characterization*, in the proceeding of the 15th International Conference on Privacy, Security and Trust, PST, Calgary, Canada (2017).
10. A. G. Manzanares, Alejandro, H. Bahsi, S. Nõmm, *KronoDroid: Timebased hybrid-featured dataset for effective android malware detection and characterization*, Comp. & Sec. 110 (2021).
11. S. Mahdavifar, A. F. A. Kadir, R. Fatemi, D. Alhadidi, A. A. Ghorbani, *Dynamic Android Malware Category Classification using Semi-Supervised Deep Learning*, in the 18th IEEE International Conference on Dependable, Autonomic, and Secure Computing (DASC), 17 (2020).
12. A. H. Lashkari, A. F. A. Kadir, L. Taheri, A. A. Ghorbani, *Toward Developing a Systematic Approach to Generate Benchmark Android Malware Datasets and*

Classification, in the proceedings of the 52nd IEEE International Carnahan Conference on Security Technology (ICCST), Montreal, Quebec, Canada, (2018).

13. A. Rahali, A. H. Lashkari, G. Kaur, L. Taheri, F. Gagnon, F. Massicotte, *DIDroid: Android Malware Classification and Characterization Using Deep Image Learning*, 10th International Conference on Communication and Network Security (ICCNS2020), Tokyo, Japan, 70 (2020).
14. M. Parkour, *Contagio Malware Dump*. https://contagiodump.blogspot.com (accessed February, 05 2023).
15. Allix, Kevin, T. F. Bissyandé, J. Klein, Y. L. Traon, *Androzoo: Collecting millions of android apps for the research community,* In IEEE/ACM 13th Working Conference on Mining Software Repositories (MSR), 468 (2016).
16. Zhou, Yajin, X. Jiang, *Dissecting android malware: Characterization and evolution*, in IEEE symposium on security and privacy, 95 (2012).
17. Maiorca, Davide, D. Ariu, I. Corona, M. Aresu, G. Giacinto, *Stealth attacks: An extended insight into the obfuscation effects on android malware*, Comp. & Sec. 51, 16 (2015).
18. Irolla, Paul, A. Dey, *The duplication issue within the drebin dataset*, Jour. of Comp. Vir.. and Hac. Tech. **14**, 245 (2018).
19. Rafiq, Husnain, N. Aslam, M. Aleem, B. Issac, R. H. Randhawa, *AndroMalPack: enhancing the ML-based malware classification by detection and removal of repacked apps for Android systems*, Sci. Rep. 12, 1 (2022).
20. *Google Play*. https://play.google.com (accessed February, 05 2023).
21. *APKPure*, https://m.apkpure.com (accessed February, 05 2023).
22. Nils Kuhnert, *VirusShare.com*. https://virusshare.com (accessed February, 05 2023).
23. *McAfee: Cyber criminals using Android malware and ransomware the most* (2013), https://www.infoworld.com/article/2614854/update--mcafee--cyber-criminals-using-android-malware-and-ransomware-the-most.html (accessed February, 05 2023).
24. *Comodo*, https://www.comodo.com/home/internet-security/security-software.php (accessed February, 05 2023).
25. L. Li, D. Li, T.F. Bissyandé, J. Klein, Y. L. Traon, D. Lo, L. Cavallaro*, Understanding android app piggybacking: a systematic study of malicious code grafting*, IEEE Trans. Inf. Forensics Sec. **12**, 1269 (2017).