MH-1M: One of The Most Comprehensive and Up-to-Date Dataset for Advanced Android Malware Detection

Hendrio Bragança[1], Vanderson Rocha[1], Joner Assolin[1], Diego Kreutz[2], Eduardo Feitosa[1]

[1]Universidade Federal do Amazonas (UFAM)
[2]LEA, PPGES — Universidade Federal do Pampa (UNIPAMPA)
{hendrio.luis, vanderson, joner.assolin, efeitosa}@icomp.ufam.br
{diegokreutz}@unipampa.edu.br

Abstract. *We introduce MH-1M, one of the most comprehensive and up-to-date dataset for advanced Android malware research. This dataset includes 1,340,515 applications, covering diverse features and extensive sets of metadata. For precise malware assessment, we utilize the VirusTotal API, integrating multiple detection methods to ensure reliable outcomes. Our GitHub repository offers users access to the processed dataset and associated metadata, totaling over 400GB. This includes comprehensive outputs from the feature extraction process and VirusTotal metadata files. Our findings underscore the important role of the MH-1M dataset as an invaluable resource for understanding the evolving landscape of malware.*

1. Introduction

The pervasive spread of Android malware poses a significant challenge for cybersecurity research. This challenge stems largely from the open-source nature and affordability of Android platforms, which grant users access to a large range market of free applications. At the same time, malware continuously evolves, adapting its tactics to execute more sophisticated and frequent attacks. Such attacks frequently result in data destruction, information theft, and several other cybercrimes [Aboaoja et al., 2022, Miranda et al., 2022, Kumar and Sharma, 2023].

Machine learning (ML) algorithms have been widely used for uncovering malware and have demonstrated remarkable effectiveness in detection systems, leveraging their discriminative capabilities to identify new variants of malicious applications [Scalas et al., 2021]. To mitigate these risks, researchers have devised diverse methods for detecting Android malware, making machine learning a key area in mobile security research [Zakeya et al., 2022].

However, the effectiveness of ML models heavily relies on the quality of the datasets used for training. Many existing datasets suffer from limitations such as outdated data, inadequate representation, and a limited number of samples and features, rendering them unsuitable for modern malware detection [Botacin et al., 2021, Miranda et al., 2022]. These issues raise concerns about the reliability of reported performance metrics and can potentially lead to misleading conclusions [Miranda et al., 2022].

For example, the Drebin-215 dataset [Yerima, 2018], released in 2018, is a subset of the original Drebin dataset developed in 2012. Therefore, models trained on it rely on outdated data that does not reflect the current landscape of malware. Similarly, the

CICInvesAndMal2019 dataset [Taheri et al., 2019], despite claiming to include data from 2019, incorporates features from outdated Android API versions (2016 and earlier).

Previous studies [Bragança et al., 2023] have shown that the diverse spectrum of Android malware activities, characterized by non-overlapping feature sets, underscores the need for multidimensional feature sets in thorough investigation and detection. While Android malware may exhibit similar behaviors and exploitation techniques, reflected in shared feature categories, the specific features used for detection can vary significantly across datasets. This complexity highlights the dynamic nature of Android malware. Additionally, recent statistics underscore a critical issue: up to 80% of AI projects fail primarily due to insufficiently representative and high-quality data in current datasets [Schmelzer, 2022, AI & Data Today, 2023].

To address this critical issue in AI projects and advance the forefront of malware detection, we introduce the MH-1M Android Malware dataset — one of the most extensive datasets ever compiled for Android malware detection. This comprehensive dataset encompasses 1,340,515 Android application packages (APKs) sourced from the Androzoo repository, spanning fourteen years from 2010 to 2024. It is worth emphasizing that one of the previous known largest dataset ever built, Drebin [Arp et al., 2014], was dated 2014 and contained 1 million samples. Following Drebin, the next most recent malware detection dataset was released in the latter half of 2023, containing roughly 100,000 samples and lacking the extensive metadata included in MH-1M.

MH-1M not only surpasses Drebin in scale but also includes updated and crucial information that is indispensable for advancing malware detection methodologies. Additionally, MH-1M provides comprehensive metadata from the extraction and analysis phases, a feature unprecedented in previous datasets. This metadata includes detailed information about the feature extraction process and analysis results, enhancing transparency and reproducibility in research efforts aimed at combating Android malware. Additionally, this rich set of structured and non-structured metadata enhances deep learning and LLM-based methods for advanced malware detection.

It is safe to say that MH-1M provides an unparalleled collection of metadata from APKs, offering remarkable insights into the evolution of malicious software spanning over more than a decade. It includes detailed Android features such as 22,394 API calls, 407 intents, 232 opcodes, and 214 permissions. These are supplemented by extensive metadata including SHA256 hashes, file names, package names, compilation APIs, and much more. In total, we offer more than 400GB of valuable data at MH-1M dataset's GitHub repository [Bragança, H. et. al., 2024], representing the largest and most comprehensive dataset ever compiled for advancing research and development in Android malware detection.

To ensure precise labeling, we utilized the VirusTotal web API to assess the threat level of each sample through diverse detection techniques. The VirusTotal API provides a comprehensive perspective on the threat level of each application. This labeling strategy enriches the dataset and offers valuable insights into the accuracy and consistency of VirusTotal's classifications.

Our contribution is threefold. Firstly, we have created the structured MH-1M dataset, which includes 1,340,515 Android applications and integrates a diverse array of

features essential for malware detection, such as intents, permissions, opcodes, and API calls. Secondly, we provide a GitHub repository with over 400GB of supplemental material, including the most extensive collection of metadata ever compiled for a malware detection dataset. Lastly, we offer a preliminary analysis based on VirusTotal labeling, enriching our understanding of malware applications, their families, and groups. Additionally, we discuss and compare MH-1M with the MH-100K dataset.

2. The MH-1M Android Malware Dataset

The MH-1M dataset represents a comprehensive repository of information on Android applications, encompassing 1,340,515 samples. The dataset is predominantly composed of benign applications, accounting for 91.1% of the total samples (1,221,421 samples), with the remaining 8.9% comprising malware data, totaling 119,094 malicious samples.

Key metadata provided in MH-1M includes the SHA256 hash (APK's signature), file name, package name, and detailed outputs from VirusTotal analysis. Additionally, the dataset features 407 intents, 214 permissions, 232 opcodes, and 22,394 API calls, offering a robust set of features for analysis. The samples in the dataset were randomly selected from the extensive list of Android applications available on Androzoo[1].

In short, the MH-1M dataset was developed using three tools: ADBuilder, AMGenerator and AMExplorer, created by the authors [Rocha et al., 2023]. ADBuilder and AMGenerator are equipped with modules for extensive data extraction and updated labeling data collection. AMExplorer, a newly developed tool, focuses on exploring features and metadata to enhance datasets for domain specialists.

Notably, fine-tuning AMExplorer's performance for constructing a consolidated metadata CSV file proved to be a challenging and time-consuming endeavor. With limited computing resources (e.g., 24 cores and 64GB of RAM), we iteratively optimized the code to efficiently process and store all necessary metadata. This effort was necessary to generate MH-1M's final 100GB CSV file of structured metadata.

The VirusTotal (VT) API[2] is a widely recognized and extensively used service for identifying potentially malicious files and URLs. VirusTotal integrates over 65 malware scanners, making it one of the most comprehensive and widely utilized services in this domain. Each VT API request retrieves a JSON file that includes metadata specific to a single sample. This metadata plays a crucial role in classifying the sample based on the number of scanners that identify the APK as potentially harmful. This approach ensures an accurate and detailed evaluation of the threat level associated with each sample, which is essential for various malware detection techniques.

3. Dataset Analysis

We conduct two major evaluations with the MH-1M dataset. Firstly, an assessment of the MH-1M dataset itself. Secondly, a first comparative study with our previously developed MH-100K dataset [Bragança et al., 2023].

To build an Android malware classification model, we utilize the XGBoost (Extreme Gradient Boosting) classifier, known for its exceptional performance in classifying

[1]https://androzoo.uni.lu/

[2]`https://developers.virustotal.com/reference/overview`

tabular data, as demonstrated in [Shwartz-Ziv and Armon, 2022]. Despite advancements in deep learning models for tabular data, research indicates that XGBoost consistently achieves superior performance across evaluated datasets.

We employed the holdout validation methodology with a 70-30 split and utilized standard classification evaluation metrics such as accuracy, precision, recall, f1-score, and the confusion matrix to analyze the results. Additionally, building on our previous findings [Bragança et al., 2023, Bragança et al., 2023], we recommend using 4 scanners as an optimal threshold for setting class thresholds in our malware classification task.

3.1. Malware Classification

As shown in Table 1, the performance of the XGBoost classifier on the MH-1M dataset demonstrates its efficacy in distinguishing between benign and malicious applications. Figure 1a further illustrates that the misclassification rate for benign applications (class 0) as malicious (class 1) was only 0.49%, indicating a highly successful outcome. Moreover, the model exhibits strong malware detection capability, with a small 11.69% misclassification rate for identifying malware applications as benign.

Class	Precision	Recall	F1-Score	Support
0	0.9887	0.9951	0.9919	366364
1	0.9462	0.8831	0.9136	35791
Accuracy		0.9851		
Macro Avg	0.9674	0.9391	0.9527	402155
Weighted Avg	0.9849	0.9851	0.9849	402155

Table 1. XGboost on the MH-1M.

Class	Precision	Recall	F1-Score	Support
0	0.9884	0.9860	0.9872	27580
1	0.8741	0.8934	0.8837	3001
Accuracy		0.9769		
Macro Avg	0.9313	0.9397	0.9354	30581
Weighted Avg	0.9772	0.9769	0.9770	30581

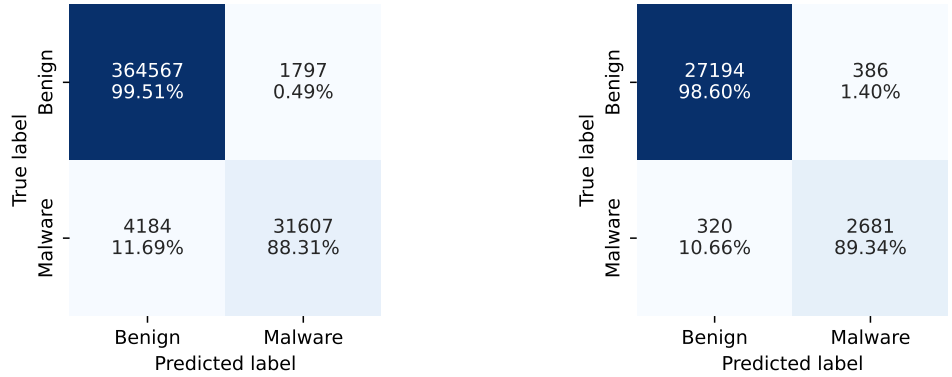
Table 2. XGboost on the MH-100K.

In the results for the MH-100K dataset, as shown in Table 2, the XGBoost classifier achieved an impressive overall accuracy of 97.69%. The precision for benign applications (class 0) was 98.84%, with a recall of 98.60% and an F1-score of 98.72%. This demonstrates the model's effectiveness in accurately identifying benign samples while maintaining a low rate of false positives.

For malicious applications (class 1), the model achieved a precision of 87.41%, recall of 89.34%, and an F1-score of 88.37%. These metrics indicate that while the model performed well overall, it faced challenges in accurately identifying all malicious samples, resulting in a slightly lower precision compared to the benign class.

For the MH-1M dataset, the XGBoost model achieved an outstanding overall accuracy of 98.51%, surpassing the results obtained with the MH-100K dataset. The precision for benign samples (class 0) was 98.87%, with a recall of 99.51% and an F1-score of 99.19%, demonstrating the model's exceptional ability to accurately identify benign samples.

Regarding malicious samples (class 1), the precision was 94.62%, recall was 88.31%, and the F1-score was 91.36%. While these metrics indicate strong performance, the slightly lower recall for malicious samples suggests that some malicious applications were misclassified as benign. Due to the MH-1M containing ten times more data than the MH-100 and having greater variability in the data, it is possible that this occurrence is a result of the dataset containing a wider range of malware.



(a) MH-1M **(b) MH-100K**

Figure 1. Confusion matrix results for XGBoot classifier.

Overall, these results underscore the XGBoost model's superior performance on the MH-1M dataset, highlighting its robustness and reliability across a broader and more diverse dataset. The increased volume of samples and extensive metadata appear crucial and contribute to improving the model's performance. However, it is worth emphasizing that researchers can further explore and utilize the more than 400GB of metadata to enhance both supervised and unsupervised machine learning models.

3.2. Cross classification and evaluation

In the following experiments, we used only shared features from both the MH-1M and MH-100K datasets. This includes a total of 250 intents, 166 permissions, 0 opcodes, and 11,545 API calls.

The analysis of cross-dataset testing results between the MH-1M and MH-100K datasets reveals significant differences in performance when employing an XGBoost model. In the first scenario, as depicted in Table 3, where the model was trained on the MH-1M dataset and tested on the MH-100K dataset, it demonstrated good results.

Class	Precision	Recall	F1-Score	Support
0	0.9929	0.9717	0.9822	92134
1	0.7783	0.9344	0.8492	9800
Accuracy		0.9681		
Macro Avg	0.8856	0.9530	0.9157	101934
Weighted Avg	0.9722	0.9681	0.9694	101934

Table 3. Training using MH-1M.

Class	Precision	Recall	F1-Score	Support
0	0.9314	0.9989	0.9640	1221421
1	0.9561	0.2450	0.3901	119094
Accuracy		0.9319		
Macro Avg	0.9437	0.6220	0.6770	1340515
Weighted Avg	0.9336	0.9319	0.9130	1340515

Table 4. Training using MH-100K.

The results showed in Figure 2a demonstrates a high degree of balance, with only 2.83% of benign applications being misclassified as malware. In contrast, a slightly higher 6.56% of malicious applications are categorized as benign. In contrast, the second scenario, where the model was trained on the MH-100K dataset and tested on the MH-1M dataset, yielded less favorable results. Specifically, the recall macro average was only 62.20%, compared to 95.30% in the former case (see Table 3).

Indeed, for the malicious class, the model's performance was suboptimal, with 75.50% of malware applications misclassified, as shown in Figure 2b. This underscores

the model's challenge in accurately classifying malicious samples from the larger and more diverse MH-1M dataset.

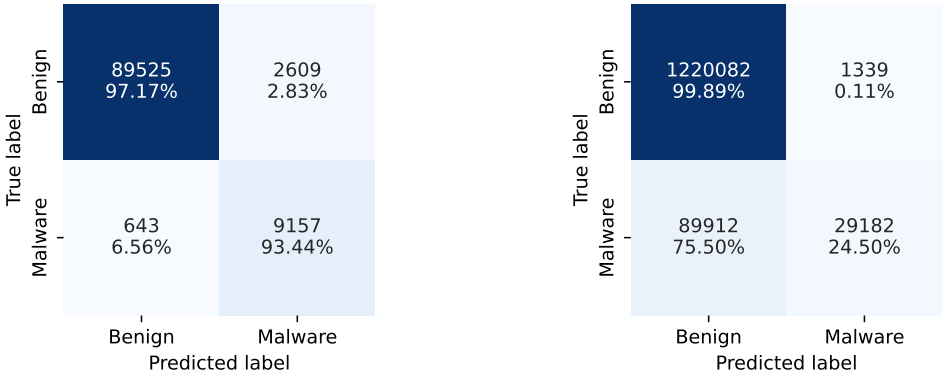


Figure 2. Confusion matrix results for cross classification.

The disparity in performance between the two cross-dataset tests can be attributed to several factors. Training the XGBoost model on the extensive and diverse MH-1M dataset provides it with a significant advantage in effectively representing both benign and malicious samples. This advantage enhances its ability to generalize and accurately classify samples from the MH-100K dataset. Conversely, when the model is trained on the smaller MH-100K dataset, it may not encounter enough exposure to the wide array of complex malware samples present in the MH-1M dataset. Consequently, its performance may diminish when tested on the MH-1M dataset. This underscores the importance of utilizing comprehensive and varied datasets for training machine learning models in the domain of malware detection.

4. Conclusions

We introduced the MH-1M dataset, a publicly available repository on GitHub [Bragança, H. et. al., 2024] comprising over 400GB of rich metadata. Inspired by the MH-100K dataset [Bragança et al., 2023], we assert that the MH-1M represents the most current, up-to-date and comprehensive dataset available, offering diverse and detailed information crucial for advancing research in Android malware detection techniques.

The MH-1M dataset offers a more comprehensive representation of both benign and malicious samples, thereby enriching the model's learning process and improving its ability to generalize across different types of malware. In conclusion, our research charts a promising path towards developing Android malware detection systems that are well-suited for real-world applications.

Agradecimentos. Esta pesquisa foi parcialmente financiada, conforme previsto nos Arts. 21 e 22 do Decreto No. 10.521/2020, nos termos da Lei Federal No. 8.387/1991, através do convênio No. 003/2021, firmado entre ICOMP/UFAM, Flextronics da Amazônia Ltda e Motorola Mobility Comércio de Produtos Eletrônicos Ltda. O presente trabalho foi realizado também com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Código de Financiamento 001 e da FAPERGS, através dos editais 08/2023 e 09/2023.

References

Aboaoja, F. A., Zainal, A., Ghaleb, F. A., Al-rimy, B. A. S., Eisa, T. A. E., and Elnour, A. A. H. (2022). Malware detection issues, challenges, and future directions: A survey. *Applied Sciences*, 12(17):8482.

AI & Data Today (2023). Top 10 reasons why ai projects fail. `https://www.aidatatoday.com/top-10-reasons-why-ai-projects-fail`.

Arp, D., Spreitzenbarth, M., Hubner, M., Gascon, H., Rieck, K., and Siemens, C. (2014). Drebin: Effective and explainable detection of android malware in your pocket. In *NDSS*, volume 14.

Botacin, M., Ceschin, F., Sun, R., Oliveira, D., and Grégio, A. (2021). Challenges and pitfalls in malware research. *Computers & Security*, 106:102287.

Bragança, H., Rocha, V., Barcellos, L. V., Souto, E., Kreutz, D., and Feitosa, E. (2023). Capturing the Behavior of Android Malware with MH-100K: A Novel and Multidimensional Dataset. In *Anais do XXIII SBSeg*, pages 510–515. SBC.

Bragança, H., Rocha, V., Souto, E., Kreutz, D., and Feitosa, E. (2023). Explaining the Effectiveness of Machine Learning in Malware Detection: Insights from Explainable AI. In *Anais do XXIII SBSeg*, Porto Alegre, RS, Brasil. SBC.

Bragança, H. et. al. (2024). MH-1M. `https://github.com/Malware-Hunter/MH-1M`.

Kumar, A. and Sharma, I. (2023). Understanding the behaviour of android ransomware attacks with real smartphones dataset. In *ICONAT*, pages 1–5. IEEE.

Miranda, T. C., Gimenez, P.-F., Lalande, J.-F., Tong, V. V. T., and Wilke, P. (2022). Debiasing android malware datasets: How can i trust your results if your dataset is biased? *IEEE Transactions on Information Forensics and Security*, 17:2182–2197.

Rocha, V., Assolin, J., Bragança, H., Kreutz, D., and Feitosa, E. (2023). AMGenerator e AM-Explorer: Geração de Metadados e Construção de Datasets Android. In *Anais Estendidos do XXIII SBSeg*, pages 41–48, Porto Alegre, RS, Brasil. SBC.

Scalas, M. et al. (2021). Malware analysis and detection with explainable machine learning. *UNICA Institutional Research Information System*.

Schmelzer, R. (2022). The one practice that is separating the AI successes from the failures. Forbes. `https://www.forbes.com/sites/cognitiveworld/2022/08/14/the-one-practice-that-is-separating-the-ai-successes-from-the-failures/`.

Shwartz-Ziv, R. and Armon, A. (2022). Tabular data: Deep learning is not all you need. *Information Fusion*, 81:84–90.

Taheri, L., Abdulkadir, A. F., and Lashkari, A. H. (2019). Investigation of the android malware (cic-invesandmal2019). `https://www.unb.ca/cic/datasets/invesandmal2019.html`.

Yerima, S. (2018). Android malware dataset for machine learning 2. `https://figshare.com/articles/dataset/Android_malware_dataset_for_machine_learning_2/5854653`.

Zakeya, N., Ségla, K., Chamseddine, T., and Alvine, B. B. (2022). Probing androvul dataset for studies on android malware classification. *Journal of King Saud University-Computer and Information Sciences*, 34(9):6883–6894.