Tainá Oliveira Soares

Detecção de Malwares Android: uma análise ampla de datasets e reprodutibilidade

Alegrete

2022

Tainá Oliveira Soares

Detecção de Malwares Android: uma análise ampla de datasets e reprodutibilidade

Trabalho de Conclusão de Curso apresentado ao Curso de Graduação em Ciência da Computação da Universidade Federal do Pampa como requisito parcial para a obtenção do título de Bacharel em Ciência da Computação.

Universidade Federal do Pampa

Orientador: Prof. Dr. Diego Kreutz

Coorientador: Prof. Dr. Eduardo Feitosa

Alegrete

2022

Ficha catalográfica elaborada automaticamente com os dados fornecidos
pelo(a) autor(a) através do Módulo de Biblioteca do
Sistema GURI (Gestão Unificada de Recursos Institucionais) .

S676d Soares, Tainá
Detecção de Malwares Android: uma análise ampla de datasets e reprodutibilidade / Tainá Soares.
36 p.

Trabalho de Conclusão de Curso(Graduação)-- Universidade Federal do Pampa, CIÊNCIA DA COMPUTAÇÃO, 2022.
"Orientação: Diego Kreutz".

1. Detecção de Malwares Android. 2. Reprodutibilidade. 3. Datasets. I. Título.

Tainá Oliveira Soares

Detecção de Malwares Android: uma análise ampla de datasets e reprodutibilidade

Trabalho de Conclusão apresentado ao Curso de Ciência da Computação da Universidade Federal do Pampa, como requisito parcial para obtenção do Título de Bacharel em Ciência da Computação.

Trabalho de Conclusão de Curso defendido e aprovado em: 11 de março de 2022.

Banca examinadora:

Prof. Dr. Diego Kreutz
Presidente da banca examinadora
Unipampa

Prof. Dr. Eduardo Luzeiro Feitosa
Coorientador
UFAM

Prof. Dr. Rodrigo Brandao Mansilha
Examinador
Unipampa

Esp. Vanderson da Silva Rocha
Examinador
UFAM

Assinado eletronicamente por **DIEGO LUIS KREUTZ**, **PROFESSOR DO MAGISTERIO SUPERIOR**, em 18/03/2022, às 12:03, conforme horário oficial de Brasília, de acordo com as normativas legais aplicáveis.

Assinado eletronicamente por **RODRIGO BRANDAO MANSILHA**, **PROFESSOR DO MAGISTERIO SUPERIOR**, em 18/03/2022, às 12:07, conforme horário oficial de Brasília, de acordo com as normativas legais aplicáveis.

Assinado eletronicamente por **Vanderson da Silva Rocha**, **Usuário Externo**, em 18/03/2022, às 12:44, conforme horário oficial de Brasília, de acordo com as normativas legais aplicáveis.

Assinado eletronicamente por **Eduardo Luzeiro Feitosa**, **Usuário Externo**, em 18/03/2022, às 15:33, conforme horário oficial de Brasília, de acordo com as normativas legais aplicáveis.

A autenticidade deste documento pode ser conferida no site https://sei.unipampa.edu.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0, informando o código verificador **0758203** e o código CRC **79B14B94**.

Este trabalho é dedicado a todos que
acreditaram no meu potencial desde
o início do curso.

AGRADECIMENTOS

Agradeço a minha mãe Tânia e as minhas irmãs Daiana e Glaucia por todo o suporte e por serem minhas maiores inspirações desde sempre. Aos tão especiais Lucas, Ícaro, Wagner e André, que sempre me ajudaram nos momentos que precisei. Ao meu sobrinho Miguel, que nasceu na última semana de elaboração desse trabalho e trouxe uma enorme onda de sentimentos bons justamente nesse período.

Agradeço a todos professores e colegas que contribuíram com a minha formação. Em especial Diego Kreutz e Eduardo Feitosa, que acreditaram na minha capacidade, orientaram e incentivaram durante toda a realização desse trabalho.

Por fim e não menos importante, agradeço aos meus colegas do projeto Malware-Hunter, que ajudaram em tudo que foi necessário durante as pesquisas.

RESUMO

Neste trabalho apresentamos resultados e discussões de pesquisa sobre os *datasets* e a reprodutibilidade dos modelos de predição no contexto de detecção de *malwares* Android. O desenvolvimento e os resultados de pesquisa estão organizados na forma de três artigos técnico-científicos, sendo dois publicados no VI Workshop Regional de Segurança da Informação e de Sistemas Computacionais (WRSeg 2021) e um terceiro em estágio avançado de desenvolvimento. Como principais contribuições pode-se destacar: (i) levantamento, análise e discussão sobre os *datasets* utilizados por pesquisas de detecção de *malwares* Android; (ii) mapeamento da disponibilidade dos *datasets*; (iii) identificação de incompletude e inconsistências nos trabalhos; (iv) catalogação e classificação de 84 fontes de dados; (v) identificação de inconsistências relacionadas à informação sobre a atualidade dos *datasets*; (vi) recomendações de boas práticas para trabalhos de pesquisa que utilizam algoritmos de aprendizado de máquina para a detecção de *malwares* Android; e (vii) identificação de obstáculos à reprodutibilidade de trabalhos.

Palavras-chave: Android. Malwares. Reprodutiblidade. Datasets.

ABSTRACT

In this work, we present an analysis and research results on datasets and the reproducibility of prediction models in the context of Android malware detection. The development and results of the research are organized in the form of three technical-scientific articles, two of which are published in the VI Workshop Regional de Segurança da Informação e de Sistemas Computacionais (WRSeg 2021) and a third stage in an advanced stage of development. As main contributions we can highlight: (i) analysis and discussion of the datasets used by researches in detection of Android malware; (ii) mapping of the datasets availability; (iii) identification of incompleteness and inconsistencies in the works; (iv) cataloging and classification of 84 data sources; (v) identification of inconsistency related to information about the actuality of the datasets; (vi) good practices recommendations for research works that use machine learning equipment to detect Android malware; and (vii) obstacles identification related to works reproducibility.

Key-words: Android. Malwares. Reproducibility. Datasets.

SUMÁRIO

1 INTRODUÇÃO

O sistema Android ocupa hoje a maior fatia de mercado de dispositivos móveis, como *smartphones* e *tablets*. Essa popularidade o torna alvo de aplicações maliciosas, que vêm crescendo rapidamente ao passar do tempo em número e sofisticação. Da mesma forma, pode-se encontrar na literatura um número crescente de trabalhos de pesquisa voltados para a detecção de *malwares* em aplicativos Android.

Os modelos de detecção de *malwares* Android mais utilizados na prática e na literatura são os baseados em aprendizado de máquina. Esses modelos classificam cada aplicativo Android, empacotado como APK (Android *Application Pack*), de acordo com premissas aprendidas durante a fase de treinamento, que acontece a partir de características de aplicativos organizadas como um conjunto estruturado de dados, conhecido como *dataset*. Consequentemente, o *dataset* é um elemento essencial e de grande impacto nos modelos detecção de *malwares*.

Nesse contexto, pode-se destacar como contribuições deste trabalho:

1. realização de um levantamento sobre o detalhamento dos *datasets* por parte das pesquisas da área de detecção de *malwares* Android;
2. mapeamento da disponibilidade dos *datasets*;
3. identificação de incompletude e inconsistências nos trabalhos;
4. identificação de inconsistências em informações sobre a atualização das fontes de dados;
5. recomendações de boas práticas para trabalhos de pesquisa que utilizam algoritmos de aprendizado de máquina;
6. catalogação e classificação de 84 fontes de dados;
7. identificação de obstáculos à reprodutibilidade de trabalhos de detecção de *malwares* Android.

Os resultados e as contribuições do trabalho estão detalhadas em dois artigos publicados no VI Workshop Regional de Segurança da Informação e de Sistemas Computacionais (WRSeg 2021).

- "Detecção de Malwares Android: *datasets* e reprodutibilidade"[1]. O artigo apresenta uma avaliação da reprodutibilidade dos *datasets* de mais de trinta trabalhos de

[1] <https://sol.sbc.org.br/index.php/errc/article/view/18540>

detecção de *malwares* Android. Uma cópia do trabalho está disponível no Apêndice A.

- "Detecção de Malwares Android: Levantamento Empírico da Disponibilidade e da Atualização das Fontes de Dados"[2]. O artigo detalha a avaliação de fontes de dados (i.e., lojas de aplicativos, *datasets* e repositórios de APKs) tipicamente utilizadas na construção de *datasets* para modelos preditivos de detecção de *malwares* Android. Uma cópia do trabalho está disponível no Apêndice B.

Finalmente, há também um trabalho em andamento, que resultará em uma terceira publicação, onde é realizada uma análise abrangente sobre trabalhos de detecção de *malwares* Android. Como contribuições desse trabalho podemos destacar:

1. Avaliação de 100 artigos de pesquisa, organizados em quatro grupos, sobre detecção de *malwares* Android;
    - Grupo 1: trabalhos citados por algum *survey* ou revisão sistemática de literatura específica do tema;
    - Grupo 2: trabalhos com 40 (ou mais) citações segundo o Google Scholar[3];
    - Grupo 3: trabalhos publicados nos principais periódicos ou conferências da área de segurança da informação;
    - Grupo 4: trabalhos publicados em periódicos ou conferências específicas da área de inteligência artificial.
2. investigação da genealogia das fontes de dados utilizadas na construção dos *datasets*;
3. identificação de problemas de reprodutibilidade nos trabalhos;
4. recomendações para pesquisas em detecção de *malwares* Android.

[2] <https://sol.sbc.org.br/index.php/errc/article/view/18541>
[3] <https://scholar.google.com>

Apêndices

APÊNDICE A – DETECÇÃO DE MALWARES ANDROID: DATASETS E REPRODUTIBILIDADE

Detecção de Malwares Android: *datasets* e reprodutibilidade

**Taina Soares[1], Guilherme Siqueira[1], Lucas Barcellos[1], Renato Sayyed[1]
Luciano Vargas[1], Gustavo Rodrigues[1], Joner Assolin[1], Jonas Pontes[2],
Eduardo Feitosa[2], Diego Kreutz[1]**

[1] Universidade Federal do Pampa (Unipampa)
[2]Universidade Federal do Amazonas (UFAM)

{{NomeSobrenome}.aluno,diegokreutz}@unipampa.edu.br
{pontes,efeitosa}@icomp.ufam.edu.br

Resumo. *Neste trabalho nós avaliamos uma amostra inicial de 38 trabalhos de pesquisa que utilizam aprendizado de máquina para detecção de malwares Android. Analisamos, em particular, o detalhamento e a disponibilidade dos datasets, que são cruciais para a validação e a reprodutibilidade do trabalho. Nossos resultados sugerem que 100% das pesquisas não são reprodutíveis por falta de informações e/ou acesso aos dados originais da pesquisa.*

1. Introdução

Os modelos de aprendizado de máquina para classificar os aplicativos Android, empacotados como APKs, entre malignos e benignos são os mais utilizados na literatura e na prática [Arslan et al., 2019]. Um modelo preditivo classifica os aplicativos de acordo com premissas que aprendeu durante a fase de treinamento, que ocorre através das características dos aplicativos organizadas como um conjunto estruturado de dados, conhecido como *dataset*. Consequentemente, a apresentação detalhada e a disponibilidade do *dataset* é imprescindível para a validação e a reprodução de trabalhos de detecção de *malwares* [Kouliaridis et al., 2020].

Neste trabalho, o objetivo é avaliarmos a reprodutibilidade, com base nos *datasets* utilizados, de estudos que propõem métodos de aprendizado de máquina para a detecção de *malwares* Android. Para alcançá-lo, coletamos 38 trabalhos existentes na literatura e realizamos um levantamento sobre a disponibilidade e o nível de detalhamento dos *datasets*.

Como contribuições deste trabalho podemos destacar: (i) realização de um levantamento inicial sobre o detalhamento dos *datasets*; (ii) mapeamento detalhado da disponibilidade dos *datasets*; (iii) identificação de incompletude e inconsistências nos trabalhos; (iv) recomendações de boas práticas para trabalhos de pesquisa que utilizem métodos de aprendizado de máquina.

O restante do trabalho está organizado da seguinte forma. Nas Seções 2 e 3 apresentamos e discutimos o levantamento de dados dos 38 trabalhos analisados. Na Seção 4 apresentamos recomendações e as considerações finais. É importante também destacar que apresentamos dados e detalhamentos adicionais na versão estendida do trabalho [Soares et al., 2021], incluindo observações empíricas gerais e o detalhamento das amostras dos *datasets* de cada um dos 38 trabalhos analisados.

2. Metodologia

Para realizar este estudo, selecionamos artigos de diferentes fontes, classificados em quatro grupos: *Grupo 1 (G1)* contém os trabalhos citados por algum *survey* ou revisão sistemática de literatura específica do tema; *Grupo 2 (G2)* inclui trabalhos com 40 (ou mais) citações segundo o Google Scholar (`https://scholar.google.com`); *Grupo 3 (G3)* contém aqueles publicados nos principais periódicos ou conferências da área de segurança, segundo o `Guide2Research.com`; e *Grupo 4 (G4)* inclui artigos publicados em conferências específicas da área de inteligência artificial. Com este último grupo, o objetivo é verificar se existe alguma diferença qualitativa significativa em termos de descrição e disponibilidade das fontes dos *datasets* quando o trabalho é publicado nessa área específica da computação, que engloba o aprendizado de máquina.

Dos 38 trabalhos que compõem este estudo, 6 são artigos retirados de revisões sistemáticas [Sharma and Rattan, 2021, Kumars et al., 2021] (*G1*). Para o grupo *G2*, resultado de uma busca no Google Scholar por "malware detection Android machine learning", foram selecionados os 14 primeiros resultados com 40 (ou mais) citações. Por fim, para os grupos *G3* e *G4*, foram selecionados 12 trabalhos publicados nas principais conferências e periódicos da área de segurança e 6 trabalhos publicados em conferências e periódicos de inteligência artificial, respectivamente.

A análise dos 38 trabalhos ocorreu em duas etapas. Na primeira, cada artigo foi analisado por dois ou três co-autores (revisores). Na segunda etapa, os artigos que resultaram em análises divergentes na primeira etapa foram novamente verificados, desta vez por um, dois ou três revisores diferentes de acordo com a complexidade das divergências. A análise de cada artigo foi guiada pelas seguintes questões: (a) Qual(is) a(s) fonte(s) de dados utilizada(s) na construção do *dataset*?; (b) A fonte de dados, que serviu como origem para os dados, é acessível? Se sim, de qual forma?; (c) Quais informações específicas (*e.g.*, quantidade, nomes, versões) sobre as aplicações Android que compõem o *dataset* são mencionadas no trabalho?

3. Resultados e Discussão

A Tabela 1 resume as informações de origem e disponibilidade dos dados dos *datasets* dos trabalhos analisados. A *Informação da origem* simplesmente registra a menção da origem dos dados nos trabalhos analisados, isto é, se o trabalho informou as fontes das quais retirou todos os dados que utilizou, definimos a coluna como *Sim*. Se apenas parte das fontes dos dados (*e.g.* de aplicações maliciosas ou benignas) foi informada, definimos como *Parcial*. E se o trabalho não informou qualquer origem dos dados, definimos como *Não*.

3.1. Detalhamento dos *datasets*

Durante a análise dos trabalhos, um dos objetivos foi identificar o nível de detalhamento da descrição dos *datasets* utilizados, mais especificamente a existência ou a ausência de informações como: (a) referência à origem das amostras utilizadas, sejam estas oriundas de um *dataset* existente ou extraídas de APKs disponíveis em um repositório; (b) detalhamento da quantidade de amostras utilizadas em cada experimento realizado; e (c) descrição da forma como o conjunto de dados próprio do trabalho foi criado (*e.g.*,

Tabela 1. Detalhamento da origem e disponibilidade dos *datasets*

Papers	Grupo	Informação da origem	Dados disponíveis
[Zhu et al., 2018], [Ali et al., 2017],	G1	Sim	Sim
[Alazab et al., 2020]	G2		
[Pendlebury et al., 2019]	G3		
[Vinod et al., 2019], [Kabakus and Dogru, 2018]	G1	Sim	Parcial
[Yuan et al., 2016], [Mahindru and Singh, 2017], [Amos et al., 2013], [Yuan et al., 2014]	G2		
[Demontis et al., 2019], [Cen et al., 2015], [Gates et al., 2014], [Ferrante et al., 2018]	G3		
[Jung et al., 2018]	G4		
[Patel and Buddadev, 2015]	G1	Parcial	Parcial
[Arora et al., 2018]	G1	Sim	Não
[Ma et al., 2019] , [Yerima et al., 2014] , [Li et al., 2018], [Mas'ud et al., 2014] , [Narudin et al., 2016]	G2		
[Chawla et al., 2021], [Fan et al., 2017], [Chen et al., 2020], [Jordaney et al., 2017], [Li et al., 2021], [Xu et al., 2016]	G3		
[Arslan et al., 2019], [Peiravian and Zhu, 2013] [Chen et al., 2018], [Mahindru and Sangal, 2021]	G4		
[Wang et al., 2019]	G1	Parcial	Não
[Wu and Hung, 2014],	G2		
[Burguera et al., 2011]	G3		
[Shabtai et al., 2012]	G4		
[Sahs and Khan, 2012], [Zarni Aung, 2013]	G2	Não	Não

combinação de *subsets* de outros *datasets*), aplicável quando um estudo utiliza particionamentos não detalhados de outros conjuntos de dados ou desenvolve suas próprias amostras.

O item (c) representa o nível mais completo de detalhamento dos *datasets*. Para que um trabalho satisfaça esse item, ele deve fornecer, além da origem dos dados e as quantidades de amostras - itens (a) e (b), um detalhamento específico dessas amostras, como os nomes e as versões das aplicações. Apesar de existirem repositórios de APKs voltados para o desenvolvimento de métodos de detecção de *malwares*, como o AndroZoo (`https://androzoo.uni.lu`), no qual são disponibilizados os nomes dos aplicativos e os resumos criptográficos, nenhum dos trabalhos analisados - nem aqueles que utilizam *subsets* de outros *datasets*, nem aqueles que desenvolvem as próprias amostras fornece essas informações necessárias para a sua reprodutibilidade.

Observando a Tabela 1, podemos visualizar as deficiências no detalhamento dos *datasets* quanto ao item (a). Embora dados referentes aos itens (b) e (c) não estejam na tabela[1], ao levarmos em consideração os itens (a) e (b), bem como a disponibilidade das fontes de dados utilizadas, aproximadamente 90% dos estudos não detalham suficientemente a origem do conjunto de dados utilizado ou não utilizam fontes disponíveis. Do total de trabalhos analisados, apenas 4 (apontados nas três primeiras linhas da tabela) mencionam a origem dos dados, utilizam fontes disponíveis e informam a quantidade de

[1]A inclusão dos itens (b) e (c) na tabela inviabilizaria o agrupamento dos trabalhos. Ao consideramos também a limitação de espaço, optamos por não representar estes itens na tabela.

amostras benignas e de *malwares* que compõem os *datasets*.

Em 12 trabalhos (aproximadamente 32%), a informação faltante é referente à quantidade de aplicativos (item b), utilizados no *dataset*, que são oriundos de lojas de aplicativos (*e.g.*, Google Play Store, AppChina, Mumayi, Amazon Appstore) ou *datasets* (*e.g.*, The Drebin Dataset, DroidKin, ContagioDump). A informação referente ao item (b) pode ser vista na tabela detalhada da versão estendida do trabalho [Soares et al., 2021]. Por exemplo, há trabalhos, como [Alazab et al., 2020], que informam a origem dos dados, mas não identificam a quantidade e nem o nome (ou resumo criptográfico) dos aplicativos retirados de cada fonte de dados. Além disso, trabalhos como [Sahs and Khan, 2012, Zarni Aung, 2013] informam o número de amostras e a distribuição do total delas em cada classe (*i.e.*, maligno ou benigno), mas não especificam a origem dos dados.

3.2. Origem dos dados

Em 60% dos trabalhos, a origem dos aplicativos benignos são lojas online de aplicativos (*e.g.*, Google Play Store, Chinese Market, Amazon Appstore App For Android, APKPure App)[2]. Entretanto, para a reconstrução do *dataset*, seriam necessárias informações como o nome e a versão dos aplicativos retirados dessas lojas. Infelizmente, nenhum dos trabalhos fornece esses detalhes.

3.3. Disponibilidade da fonte dos dados

Dos trabalhos analisados e que mencionam pelo menos alguma origem de dados, apenas quatro ([Pendlebury et al., 2019], [Ali et al., 2017], [Zhu et al., 2018], [Alazab et al., 2020]) possuem todas as origens disponíveis. As fontes de dados citadas por estes são AndroZoo, ContagioDump, MalShare, VirusShare e M0Droid[3].

Em aproximadamente 58% dos trabalhos, aqueles em que, na Tabela 1, a coluna *Dados disponíveis* está como *Não*, as fontes referenciadas são inacessíveis, como é o caso de trabalhos como [Jordaney et al., 2017] e [Chawla et al., 2021]. É interessante destacarmos também que alguns trabalhos, como [Shabtai et al., 2012], relatam que as amostras utilizadas no experimento foram desenvolvidas internamente, porém sem fornecer detalhes ou o acesso à tais amostras. Em todos esses casos, temos problemas que afetam a reprodutibilidade dos trabalhos, como é evidente.

4. Considerações Finais

A partir da análise minuciosa de 38 *papers*, podemos concluir que todos os trabalhos falham em apresentar pelo menos alguma informação fundamental acerca dos *datasets* (*e.g.*, origem dos dados, quantidade de aplicativos) ou não indicam a forma de acessar a fonte de dados utilizada na construção do *dataset*. Resumidamente, podemos assumir que os dados coletados indicam que a maioria das pesquisas em detecção de *malwares* Android não são reprodutíveis e nem verificáveis devido à falta de informação sobre os dados utilizados. Esse cenário traz impactos negativos, por exemplo, na construção de

[2]`https://play.google.com/store`, `https://shouji.baidu.com/`, `https://www.amazon.com/gp/mas/get/amazonapp`, `https://m.apkpure.com`

[3]`http://contagiominidump.blogspot.com/`, `https://malshare.com/`, `https://virusshare.com/`, `https://www.azsecure-data.org/other-data.html`

novos modelos de aprendizado de máquina, uma vez que a comparação é comprometida pela inviabilidade de reprodução dos experimentos existentes na literatura.

Como **recomendações**, destacamos que o detalhamento dos *datasets* deve incluir as fontes utilizadas, sejam estas repositórios de APKs ou *datasets* de terceiros. Além disso, é importante informar o *subset* utilizado no treinamento e validação dos modelos de aprendizado de máquina. Idealmente, recomendamos que sejam utilizadas fontes públicas para extrair as amostras, facilitando e acelerando a reprodução dos *datasets*. Complementarmente, a disponibilidade do conjunto exato de dados, utilizado no trabalho, viabilizaria uma reprodução fidedigna da pesquisa. É importante ressaltar também que devemos evitar fontes de dados antigas (*e.g.*, *datasets* de 2012 – a API do Android sofreu modificações significativas em 2015, por exemplo), pois não há garantias que os padrões encontrados pelos modelos preditivos, em amostras antigas, sejam aplicáveis em *malwares* atuais.

Dentre os **trabalhos futuros**, destacamos: analisar aspectos de reprodutibilidade dos modelos de aprendizado de máquina (*e.g.*, bibliotecas e hiperparâmetros utilizados).

Agradecimentos

Esta pesquisa foi 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.

Referências

Alazab, M., Alazab, M., Shalaginov, A., Mesleh, A., and Awajan, A. (2020). Intelligent mobile malware detection using permission requests and api calls. *Future Generation Computer Systems*, 107:509–521.

Ali, M. A., Svetinovic, D., Aung, Z., and Lukman, S. (2017). Malware detection in android mobile platform using machine learning algorithms. In *2017 International Conference on Infocom Technologies and Unmanned Systems (Trends and Future Directions) (ICTUS)*, pages 763–768.

Amos, B., Turner, H., and White, J. (2013). Applying machine learning classifiers to dynamic android malware detection at scale. In *9th International Wireless Communications and Mobile Computing Conference (IWCMC)*, pages 1666–1671.

Arora, A., Peddoju, S. K., Chouhan, V., and Chaudhary, A. (2018). Hybrid android malware detection by combining supervised and unsupervised learning. In *Proceedings of the 24th Annual International Conference on Mobile Computing and Networking*, page 798–800. ACM.

Arslan, R. S., Doğru, İ. A., and Barişçi, N. (2019). Permission-based malware detection system for android using machine learning techniques. *International journal of software engineering and knowledge engineering.*, 29(01):43–61.

Burguera, I., Zurutuza, U., and Nadjm-Tehrani, S. (2011). Crowdroid: Behavior-based malware detection system for android. In *Proceedings of the 1st ACM Workshop on Security and Privacy in Smartphones and Mobile Devices*, page 15–26. ACM.

Cen, L., Gates, C. S., Si, L., and Li, N. (2015). A probabilistic discriminative model for android malware detection with decompiled source code. *IEEE Transactions on Dependable and Secure Computing*, 12(4):400–412.

Chawla, N., Kumar, H., and Mukhopadhyay, S. (2021). Machine learning in wavelet domain for electromagnetic emission based malware analysis. *IEEE Transactions on Information Forensics and Security*, 16:3426–3441.

Chen, X., Li, C., Wang, D., Wen, S., Zhang, J., Nepal, S., Xiang, Y., and Ren, K. (2020). Android hiv: A study of repackaging malware for evading machine-learning detection. *IEEE Transactions on Information Forensics and Security*, 15:987–1001.

Chen, Z., Yan, Q., Han, H., Wang, S., Peng, L., Wang, L., and Yang, B. (2018). Machine learning based mobile malware detection using highly imbalanced network traffic. *Information Sciences*, 433-434:346–364.

Demontis, A., Melis, M., Biggio, B., Maiorca, D., Arp, D., Rieck, K., Corona, I., Giacinto, G., and Roli, F. (2019). Yes, machine learning can be more secure! a case study on android malware detection. *IEEE Transactions on Dependable and Secure Computing*, 16(4):711–724.

Fan, M., Liu, J., Wang, W., Li, H., Tian, Z., and Liu, T. (2017). Dapasa: Detecting android piggybacked apps through sensitive subgraph analysis. *IEEE Transactions on Information Forensics and Security*, 12(8):1772–1785.

Ferrante, A., Malek, M., Martinelli, F., Mercaldo, F., and Milosevic, J. (2018). Extinguishing ransomware - a hybrid approach to android ransomware detection. In Imine, A., Fernandez, J. M., Marion, J.-Y., Logrippo, L., and Garcia-Alfaro, J., editors, *Foundations and Practice of Security*, pages 242–258, Cham. Springer International Publishing.

Gates, C. S., Li, N., Peng, H., Sarma, B., Qi, Y., Potharaju, R., Nita-Rotaru, C., and Molloy, I. (2014). Generating summary risk scores for mobile applications. *IEEE Transactions on Dependable and Secure Computing*, 11(3):238–251.

Jordaney, R., Sharad, K., Dash, S. K., Wang, Z., Papini, D., Nouretdinov, I., and Cavallaro, L. (2017). Transcend: Detecting concept drift in malware classification models. In *26th USENIX Security Symposium*, pages 625–642. USENIX Association.

Jung, J., Kim, H., Shin, D., Lee, M., Lee, H., Cho, S.-j., and Suh, K. (2018). Android malware detection based on useful api calls and machine learning. In *IEEE First International Conference on Artificial Intelligence and Knowledge Engineering (AIKE)*, pages 175–178.

Kabakus, A. T. and Dogru, I. A. (2018). An in-depth analysis of android malware using hybrid techniques. *Digital Investigation*, 24:25–33.

Kouliaridis, V., Kambourakis, G., and Peng, T. (2020). Feature importance in android malware detection. In *IEEE 19th International Conference on Trust, Security and Privacy in Computing and Communications (TrustCom)*, pages 1449–1454.

Kumars, R., Alazab, M., and Wang, W. (2021). *A Survey of Intelligent Techniques for Android Malware Detection*, pages 121–162. Springer International Publishing, Cham.

Li, C., Chen, X., Wang, D., Wen, S., Ahmed, M. E., Camtepe, S., and Xiang, Y. (2021). Backdoor attack on machine learning based android malware detectors. *IEEE Transactions on Dependable and Secure Computing*, pages 1–1.

Li, J., Sun, L., Yan, Q., Li, Z., Srisa-an, W., and Ye, H. (2018). Significant permission identification for machine-learning-based android malware detection. *IEEE Transactions on Industrial Informatics*, 14(7):3216–3225.

Ma, Z., Ge, H., Liu, Y., Zhao, M., and Ma, J. (2019). A combination method for android malware detection based on control flow graphs and machine learning algorithms. *IEEE Access*, 7:21235–21245.

Mahindru, A. and Sangal, A. L. (2021). MLDroid—framework for Android malware detection using machine learning techniques. *Neural Computing and Applications*, 33(10):5183–5240.

Mahindru, A. and Singh, P. (2017). Dynamic permissions based android malware detection using machine learning techniques. In *Proceedings of the 10th Innovations in Software Engineering Conference*, page 202–210. ACM.

Mas'ud, M. Z., Sahib, S., Abdollah, M. F., Selamat, S. R., and Yusof, R. (2014). Analysis of features selection and machine learning classifier in android malware detection. In *International Conference on Information Science Applications*, pages 1–5.

Narudin, F. A., Feizollah, A., Anuar, N. B., and Gani, A. (2016). Evaluation of machine learning classifiers for mobile malware detection. *Soft Computing*, 20(1):343–357.

Patel, K. and Buddadev, B. (2015). Detection and mitigation of android malware through hybrid approach. In Abawajy, J. H., Mukherjea, S., Thampi, S. M., and Ruiz-Martínez, A., editors, *Security in Computing and Communications*, pages 455–463, Cham. Springer International Publishing.

Peiravian, N. and Zhu, X. (2013). Machine learning for android malware detection using permission and api calls. In *IEEE 25th International Conference on Tools with Artificial Intelligence*, pages 300–305.

Pendlebury, F., Pierazzi, F., Jordaney, R., Kinder, J., and Cavallaro, L. (2019). TESSERACT: Eliminating experimental bias in malware classification across space and time. In *28th USENIX Security Symposium (USENIX Security 19)*, pages 729–746, Santa Clara, CA. USENIX Association.

Sahs, J. and Khan, L. (2012). A machine learning approach to android malware detection. In *European Intelligence and Security Informatics Conference*, pages 141–147.

Shabtai, A., Kanonov, U., Elovici, Y., Glezer, C., and Weiss, Y. (2012). "Andromaly": a behavioral malware detection framework for android devices. *Journal of Intelligent Information Systems*, 38(1):161–190.

Sharma, T. and Rattan, D. (2021). Malicious application detection in android — a systematic literature review. *Computer Science Review*, 40:100373.

Soares, T., Siqueira, G., Barcellos, L., Sayyed, R., Vargas, L., Rodrigues, G., Assolin, J., Pontes, J., Feitosa, E., and Kreutz, D. (2021). Detecção de malwares android: datasets e reprodutibilidade. `https://arxiv.kreutz.xyz/wrseg2021reprodutibilidade_ve1.pdf`.

Vinod, P., Zemmari, A., and Conti, M. (2019). A machine learning based approach to detect malicious android apps using discriminant system calls. *Future Generation Computer Systems.*, 94:333–350.

Wang, S., Chen, Z., Yan, Q., Yang, B., Peng, L., and Jia, Z. (2019). A mobile malware detection method using behavior features in network traffic. *Journal of Network and Computer Applications*, 133:15–25.

Wu, W.-C. and Hung, S.-H. (2014). Droiddolphin: A dynamic android malware detection framework using big data and machine learning. In *Proceedings of the 2014 Conference on Research in Adaptive and Convergent Systems*, page 247–252. ACM.

Xu, K., Li, Y., and Deng, R. H. (2016). Iccdetector: Icc-based malware detection on android. *IEEE Transactions on Information Forensics and Security*, 11(6):1252–1264.

Yerima, S. Y., Sezer, S., and Muttik, I. (2014). Android malware detection using parallel machine learning classifiers. In *Eighth International Conference on Next Generation Mobile Apps, Services and Technologies*, pages 37–42.

Yuan, Z., Lu, Y., Wang, Z., and Xue, Y. (2014). Droid-sec: Deep learning in android malware detection. *SIGCOMM Comput. Commun. Rev.*, 44(4):371–372.

Yuan, Z., Lu, Y., and Xue, Y. (2016). Droiddetector: android malware characterization and detection using deep learning. *Tsinghua Science and Technology*, 21(1):114–123.

Zarni Aung, W. Z. (2013). Permission-based android malware detection. *International Journal of Scientific & Technology Research*, 2(3):228–234.

Zhu, H.-J., You, Z.-H., Zhu, Z.-X., Shi, W.-L., Chen, X., and Cheng, L. (2018). Droiddet: Effective and robust detection of android malware using static analysis along with rotation forest model. *Neurocomputing*, 272:638–646.

APÊNDICE B – DETECÇÃO DE MALWARES ANDROID: LEVANTAMENTO EMPÍRICO DA DISPONIBILIDADE E DA ATUALIZAÇÃO DAS FONTES DE DADOS

Detecção de Malwares Android: Levantamento Empírico da Disponibilidade e da Atualização das Fontes de Dados

Tainá Soares[1], Joner Mello[1], Lucas Barcellos[1], Renato Sayyed[1], Guilherme Siqueira[1] Karina Casola[1], Estevão Costa[2], Nicolas Gustavo[2], Eduardo Feitosa[2], Diego Kreutz[1],

[1] Universidade Federal do Pampa (Unipampa)
[2]Universidade Federal do Amazonas (UFAM)

Resumo. *Neste estudo avaliamos 84 fontes de dados utilizadas para a concepção de modelos de aprendizado de máquina aplicados à detecção de malwares Android, sendo 39 lojas de aplicativos, 30 datasets e 15 repositórios de APKs. Verificamos que 68,75% dos trabalhos utilizam fontes de dados antigas, mesmo existindo opções de fontes atuais. Também observamos que a disponibilidade e a corretude dos registros das fontes de dados nem sempre são condizentes com o informado e, consequentemente, podem impactar negativamente a qualidade dos métodos de detecção de malwares.*

1. Introdução

Para modelos de detecção de *malwares* Android, a atualidade dos *datasets* é importante e pode impactar diretamente o desempenho da solução [Allix et al., 2015]. Como os *malwares* Android possuem uma natureza dinâmica, modelos de aprendizado de máquina conseguem reconhecer um aplicativo malicioso atual somente se os dados de treino incluírem informações sobre o comportamento de *malwares* atuais.

Na prática, muitos *datasets* atuais são construídos a partir de *datasets* mais antigos [Sharma and Rattan, 2021], o que também representa um problema. Por exemplo, o *dataset* Drebin-215 (disponibilizado em 2018) é constituído, na verdade, por um subconjunto de dados do *dataset* Drebin, datado de 2012. O mesmo ocorre com diversos outros conjuntos de dados, como o Android Botnet, formado por dados do Malware Genome Project, Contagio Mini Dump e VirusTotal. Além disso, há *datasets*, como o CICInvesAndMal2019, que afirmam incluir dados atuais, de 2019, porém, contêm apenas características de versões bastante antigas (e.g., 2016 e inferior) da API do Android.

Neste trabalho temos como objetivos: (a) realizar um levantamento detalhado de informações sobre a atualização e a disponibilidade de fontes de dados; e (b) investigar as fontes de dados utilizadas na prática por trabalhos que propõem modelos de aprendizado de máquina para detecção de *malwares* Android. Como contribuições resultantes do desenvolvimento destes objetivos, podemos destacar: (a) catalogação e classificação de 84 fontes de dados com relação ao tipo, disponibilidade e atualização; (b) avaliação e discussão sobre as fontes de dados utilizadas nas pesquisas de 35 trabalhos acadêmicos; (c) identificação de inconsistências na informação de atualização das fontes de dados; e (d) identificação de obstáculos à reprodutibilidade dos trabalhos.

Em estudos similares, como o [Kouliaridis et al., 2020], os autores limitam-se ao estudo de 10 *datasets*, onde comparam características como idade, tamanho, acesso (disponível, indisponível ou através de solicitação) e destacam a presença de características como permissões e *intents* em 30% dos *datasets*. O nosso estudo cobre um número mais

expressivo de fontes de dados (84) e investiga detalhadamente questões relacionadas à disponibilidade e atualização das fontes utilizadas em estudos atuais.

O trabalho está organizado como segue. Nas Seções 2 e 3, apresentamos as 84 fontes de dados catalogadas e uma discussão sobre as fontes de dados utilizadas nos 35 trabalhos analisados, respectivamente. Finalmente, nas Seções 4 e 5, apresentamos questões sobre a atualização das fontes de dados e as considerações finais, respectivamente.

2. Atualização e Disponibilidade das Fontes de Dados

O conjunto de dados analisado neste trabalho é composto por 84 fontes, sendo 39 lojas de aplicativos Android, 30 *datasets* e 15 repositórios de Pacotes de Aplicação Android (*Android Application Pack* ou simplesmente APKs). Essas fontes foram catalogadas a partir: (a) da revisão sistemática sobre detecção de aplicações maliciosas no Android [Sharma and Rattan, 2021], (b) de 35 trabalhos selecionados para análise (conforme detalhado em [Soares et al., 2021b]) e (c) dos 100 primeiros resultados da busca por "Android Dataset" no Google Dataset Search[1], Kaggle[2] e FigShare[3]. A relação completa e detalhada das fontes está disponível na versão estendida do trabalho [Soares et al., 2021a].

As fontes foram classificadas, em relação à disponibilidade, em três tipos: disponível, indisponível e restrito. Para realizar essa classificação, foram utilizadas as duas convenções a seguir: **Fontes não localizadas** são aquelas não encontradas nos nossos processos de busca, que foram (a) buscas *web* pelo nome da fonte e (b) verificação de todos os *links* retornados das duas primeiras páginas de resultado. As buscas, por cada fonte de dados, foram realizadas por, no mínimo, dois co-autores do trabalho. **Fontes sem acesso público** são aquelas sendo encontradas através das nossas buscas, isto é, o *link* da fonte foi encontrado, mas na página (a) não há informações sobre como acessar os dados (e.g., se é necessário enviar um *e-mail*, preencher um formulário ou solicitar previamente algum tipo de autorização) ou (b) é informado que a fonte não está mais disponível.

A classificação de uma fonte como disponível significa que ela foi localizada e seu acesso é público, isto é, sem restrições (e.g., autorização prévia ou credenciais de acesso). As fontes não localizadas ou sem acesso público foram classificadas como indisponíveis, como é o caso do Malware Genome Dataset. Finalmente, todas as fontes que exigem alguma autorização prévia (e.g., contato via e-mail ou formulário) ou credenciais (e.g., login e senha) para o acesso foram classificadas como restritas. A classificação quanto a disponibilidade das fontes foi baseada no trabalho de [Kouliaridis et al., 2020], que classifica 10 *datasets*.

As fontes que requerem solicitação de acesso, mas não responderam às solicitações em 30 dias ou mais, como é o caso do The Drebin Dataset, e aquelas que exigem credenciais mediante pagamento, como é o caso da Virus Total Malware Service Intelligence, foram classificadas como indisponíveis.

2.1. Lojas de Aplicativos

As lojas (ou mercados) de aplicativos são plataformas que servem como meio de distribuição de *software* para dispositivos móveis, como *smartphones* e *tablets*, baseados

[1] `https://datasetsearch.research.google.com/`
[2] `https://www.kaggle.com/`
[3] `https://figshare.com/`

em Android. Essas lojas armazenam o APK de cada aplicativo disponibilizado por elas, que é utilizado para instalar o aplicativo nos dispositivos. O principal exemplo de mercado de aplicativos Android é a Google Play Store, loja oficial para o sistema operacional Android.

Das 39 lojas de aplicativos, 25 foram classificadas como disponíveis e destas, 84% (21 delas) são atualizadas constantemente. As lojas disponíveis são aquelas que possuem um site oficial acessível e alguma forma (e.g., *link*) para o download dos APKs, como é o caso do mercado AndroidLista. Por outro lado, há 14 lojas que classificamos como indisponíveis por não terem um meio de acesso aos aplicativos. Na maior parte dos casos, como o AndroidDrawer, há problemas no acesso do site da loja — site não abre ou retorna erro. Há casos de lojas, como a GFan, Anruan e 10086, onde os sites oficiais retornam erro de servidor não encontrado para diferentes navegadores *web* populares (e.g., Google Chrome, Mozilla Firefox) e, foram classificadas como indisponíveis.

2.2. *Datasets* e Repositórios de APKs

A classificação quanto a atualização dos *datasets* e repositórios de APKs foi realizada utilizando intervalos de tempo ([2008-2012], [2013-2017] e [2018-2021]) para agrupar os dados. As datas consideradas são aquelas informadas nos estudos que originaram as fontes ou nos sites delas. Se a data informada é 2015, como no caso do Wang's Repository, classificamos a atualização da fonte como contida no intervalo [2013-2017]. Os detalhes completos, de todos os repositórios e *datasets*, podem ser vistos em [Soares et al., 2021a].

Dos 45 *datasets* e repositórios, 25 são disponíveis, 11 são restritos e 9 são indisponíveis. Do total, 20 podem ser considerados como atualizados, isto é, estão contidos no intervalo de 2018 e 2021 (parâmetro utilizado neste estudo). Outra observação interessante é o fato de a maioria das fontes consideradas disponíveis serem também as mais atuais: 16 de um total de 25 têm atualização entre 2018 e 2021, conforme pode ser observado na Tabela 1.

Um resumo do período de atualização dos *datasets* e dos repositórios de APKs classificados como disponíveis ou restritos pode ser visto no gráfico da Figura 1. Cerca de 55% dessas fontes (conjunto das disponíveis ou restritas) têm atualização entre 2018 e 2021, ou seja, mais da metade são consideradas recentes. As informações de disponibilidade e quantidade das fontes de dados dispostas por período de atualização sugerem haver uma tendência em tornar os dados públicos e de fácil acesso.

Há alguns casos de fontes de dados disponíveis (e.g., CICMalDroid 2020) que disponibilizam tanto *datasets* quanto repositórios de APKs. Outra característica dessa fonte é que ela é composta por amostras de *malwares* coletadas de diversas outras fontes: VirusTotal, AMD, contagio, entre outras. Além do CICMalDroid 2020, outras fontes também contêm dados de diversas origens, como é o caso do Android Botnet e CIC-AndMal2017.

3. Trabalhos e Fontes de Dados

Anteriormente, analisamos a reprodutibilidade de 35 artigos científicos [Soares et al., 2021b]. Neste trabalho, realizamos uma análise das 31 fontes de dados utilizadas pelos trabalhos, das quais 11 são lojas de aplicativos e 20 são *datasets* ou repositórios de APKs.

Tabela 1. *Datasets* e Repositórios de APKs: Atualização e Disponibilidade

***Datasets* e Repositórios de APKs**	**Atualização**	**Acesso**
Ether Malware Analysis Dataset	[2008-2012]	Disponível
Contagio Malware Dump, CIC-AAGM2017, MudFlow, Android Botnet, M0Droid, GaziBenignApp, Heldroid	[2013-2017]	
Contagio Mobile, Android Permissions Dataset, *VirusShare*, **CICInvesAndMal2019**, *TheZoo*, *AndroMalShare*, Koodous, Drebin-215, Dataset of Android Permissions, **CICMalDroid 2020**, Android Malware and Normal Permissions Dataset, Android Malware and Benign Application Dataset, PARUDroid, *Comodo Cloud Security Center*, **Wang's Repository**, Drebin4000 and AMD6000	[2018-2021]	
MobileSandbox project (MobWorm)	Atualização Não Encontrada	
CIC-AndMal2017, *AndroZoo*, Andro-AutoPsy, Andro-Dumpsys, Andro-Profiler, Andro-Tracker	[2013-2017]	Restrito
UpDroid, *Contagio Mini Dump*, *COVID19 Apps*, CCCS-CIC-AndMal-2020	[2013-2018]	
DroidKin	Atualização Não Encontrada	
The Drebin Dataset, Android Malware Genome Project	[2008-2012]	Indisponível
Android PRAGuard Dataset, PlayDrone Project	[2013-2017]	
McAfee, *Inter-Component Communication (IcRE) Repository*, *New Malware Families 2015*, *Virus Total Malware Service Intelligence*, Kharon Malware Dataset	Atualização Não Encontrada	

Os repositórios de APKs estão em *itálico*. Fontes que disponibilizam tanto *datasets* quanto repositórios de APKs estão em negrito.

3.1. Lojas de Aplicativos

As lojas de aplicativos são comumente utilizadas como fonte de aplicativos benignos para compor os *datasets* [Wang et al., 2019]. Em 21 dos 35 trabalhos analisados, os *datasets* têm dados oriundos de lojas, como Google Play Store, SlideME e PandaApp. Do total de 11 lojas, 8 foram classificadas como disponíveis e todas, exceto a SlideME (`http://slideme.org/`), são atualizadas constantemente. A disponibilidade detalhada dessas e das demais lojas pode ser vista em [Soares et al., 2021a]. A SlideME incluiu um aplicativo novo pela última vez em 2019, mas em 2021 atualizou um aplicativo da sua loja.

A disponibilidade e atualização da maioria das lojas pode sugerir que os dados utilizados pelos trabalhos são atuais e de fácil acesso. Infelizmente, todos os trabalhos falham em relação ao acesso, pois não fornecem os dados necessários sobre os aplicativos, como os nomes e versões, que são fundamentais para a reprodutibilidade dos trabalhos.

3.2. *Datasets* e Repositórios de APKs

Dos 20 *datasets* e repositórios de APKs, 45% foram classificados como disponíveis, 40% como indisponíveis e 15% como acesso restrito. É interessante observar que 65% dos trabalhos utilizam pelo menos uma das 8 fontes indisponíveis.

Considerando os 16 trabalhos mais recentes, publicados de 2018 a 2021, podemos verificar a atualização e a disponibilidade das fontes de dados na Figura 1. Como pode ser observado, 12 dos 16 trabalhos citam alguma fonte de dados atual (em relação aos seus respectivos anos de publicação) disponível ou restrita. Entretanto, todos os trabalhos utilizam pelo menos uma fonte de dados com atualização entre 2012 e 2016. Consequentemente, podemos concluir que nenhum desses trabalhos utiliza um conjunto de dados atualizado, pois todos os *datasets* contêm também dados de fontes defasadas, i.e., antigas.

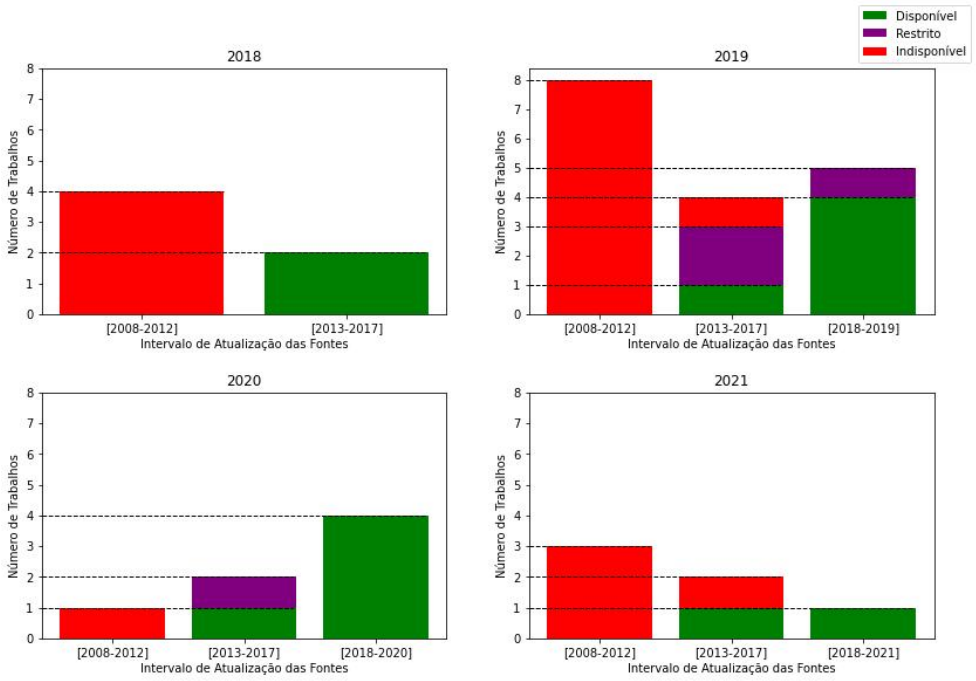

Figura 1. Atualização e Disponibilidade das Fontes de Dados Utilizadas

4. Atualização das Fontes de Dados

Um ponto a ser analisado acerca das datas das fontes é a possibilidade de diferença entre as datas informadas nos trabalhos ou *sites* e as datas das APIs presentes em determinada fonte de dados. Esse é o caso das fontes CIC-InvesAndMal2019[4] e CICMalDroid2020[5].

As fontes de dados CIC-InvesAndMal2019 e CICMalDroid2020 foram analisadas quanto às versões das APIs da seguinte maneira: (i) primeiramente, o *download* dos APKs de cada uma delas foi realizado; (ii) depois disso, um algoritmo que seleciona os APKs válidos foi executado sobre os dados, pois há alguns arquivos APK corrompidos (*i.e.* mal formados, impossíveis de serem analisados); (iii) por fim, foram verificadas as APIs presentes nos arquivos válidos.

O CIC-InvesAndMal2019 teve os dados coletados de 2012 a 2019 e, portanto, tem 2019 como ano de atualização. A questão é que essa fonte contém APIs de versões de antes de 2012 e não contém nenhuma API de 2019. Ou seja, o CIC-InvesAndMal2019 é considerada uma fonte de 2019, mas não contém dados de APIs de 2019 (a mais recente é de 2016). Já no caso do CICMalDroid2020, foi verificada a presença de dados de aplicações de APIs atuais, mas a grande maioria destas são aplicações benignas. Este também é um problema, pois por mais que seja verdade que a fonte contenha dados de APIs atuais, ela continua sendo desatualizada quanto às aplicações malignas, que são a parte mais importante quando o contexto é o uso desses dados para detecção de *malwares* e aprendizado de máquina.

[4]`https://www.unb.ca/cic/datasets/invesandmal2019.html`
[5]`https://www.unb.ca/cic/datasets/maldroid-2020.html`

5. Considerações Finais

As principais conclusões do nosso estudo podem ser separadas de acordo com as duas análises realizadas: (a) panorama das 84 fontes de dados e (b) uso das fontes de dados por parte dos 35 trabalhos. Com relação à análise (a), onde foram considerados aspectos de atualização e disponibilidade de 84 fontes de dados, as principais conclusões são: (i) a maioria (59,52%) das fontes de dados são disponíveis (aproximadamente 55% dos *datasets* e repositórios de APKs e 64% dos mercados de aplicativos Android); e (ii) existe uma tendência em tornar as fontes de dados disponíveis, pois quanto mais atual o período de tempo analisado, maior a quantidade de fontes disponíveis.

Já quanto a análise (b), podemos concluir que as fontes de dados desatualizadas e indisponíveis ainda são bastante utilizadas em pesquisas atuais da área. Esse fato é um problema, pois um modelo de aprendizado de máquina não pode garantir a identificação de uma nova linhagem de *malware* que não é representada no conjunto de dados de treinamento [Allix et al., 2015]. A atualização regular de *datasets* e a proximidade histórica destes conjuntos de dados são fundamentais para a minimizar as ameaças à validade destes estudos.

Como trabalhos futuros podemos destacar: (a) analisar as versões das APIs presentes nas fontes de dados para validar a atualização; (b) realizar o mapeamento de relação de origem entre as fontes de dados para verificar o quão novas as fontes são; (c) avaliar o impacto de *datasets* de diferentes períodos nos modelos de detecção de *malwares*; e (d) analisar as fontes de dados de acordo com a possibilidade de acesso às versões antigas das aplicações.

Agradecimentos

Esta pesquisa foi 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.

Referências

Allix, K., Bissyandé, T. F., Klein, J., and Le Traon, Y. (2015). Are your training datasets yet relevant? In Piessens, F., Caballero, J., and Bielova, N., editors, *Engineering Secure Software and Systems*, pages 51–67, Cham. Springer International Publishing.

Kouliaridis, V., Kambourakis, G., and Peng, T. (2020). Feature importance in mobile malware detection. *CoRR*, abs/2008.05299.

Sharma, T. and Rattan, D. (2021). Malicious application detection in android — a systematic literature review. *Computer Science Review*, 40:100373.

Soares, T., Mello, J., Barcellos, L., Sayyed, R., Siqueira, G., Casola, K., Costa, E., Gustavo, N., Feitosa, E., and Kreutz, D. (2021a). Detecção de malwares android: Levantamento empírico da disponibilidade e da atualização das fontes de dados (versão estendida). `https://arxiv.kreutz.xyz/wrseg2021_disponibilidade_ve1.pdf`.

Soares, T., Siqueira, G., Barcellos, L., Sayyed, R., Vargas, L., Rodrigues, G., Assolin, J., Pontes, J., and Kreutz, D. (2021b). Detecção de malwares android: datasets e reprodutibilidade. `https://arxiv.kreutz.xyz/mh21_reprodutibilidade.pdf`.

Wang, S., Chen, Z., Yan, Q., Yang, B., Peng, L., and Jia, Z. (2019). A mobile malware detection method using behavior features in network traffic. *Journal of Network and Computer Applications*, 133:15–25.