Detecção de Malwares Android: Levantamento Empírico da Disponibilidade e da Atualização das Fontes de Dados (versão estendida)*

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

[1]Laboratório de Estudos Avançados em Computação (LEA)
Programa de Pós-Graduação em Engenharia de Software (PPGES)
Universidade Federal do Pampa (Unipampa)

[2]Grupo de Pesquisa em Tecnologias Emergentes e Segurança de Sistemas (ETSS)
Instituto de Computação (IComp)
Universidade Federal do Amazonas (UFAM)

{NomeSobrenome}.aluno@unipampa.edu.br
{ecc,nicolas.neves,vanderson,efeitosa}@icomp.ufam.edu.br
diegokreutz@unipampa.edu.br

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

Diversos trabalhos recentes vêm explorando técnicas e modelos de aprendizado de máquina para identificar e combater a proliferação de *malwares* em sistemas Android [Yan and Yan, 2018, Wang et al., 2019, Arslan et al., 2019, Ming et al., 2020], mais especificamente no âmbito dos modelos supervisionados de classificação. Um elemento fundamental para este tipo de detecção é o *dataset*, que é aplicado no treinamento, no teste e na validação dos modelos propostos. É através dele que os modelos aprendem o que considerar para atingir o objetivo de classificar um aplicativo Android como *malware* ou não. Consequentemente, o desempenho dos modelos possui uma forte relação com o *dataset* utilizado [Allix et al., 2015].

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, Wei et al., 2017]. Como os *malwares* estão sempre em evolução, modelos de aprendizado de máquina conseguem reconhecer um aplicativo malicioso atual somente se os dados de treino incluírem informações sobre o comportamento atual desses *malwares*. Portanto, *datasets* atuais são especialmente importantes em um cenário onde há

*Este artigo é uma versão estendida do paper [Soares et al., 2021a], de 6 páginas, originalmente publicado no WRSeg 2021 e convidado para publicação na ReABTIC.

um constante desenvolvimento e uma crescente sofisticação de novos *malwares* Android [SophosLabs, 2021].

Além da atualidade, a disponibilidade dos *datasets* também é importante, pois permite a reprodutibilidade dos experimentos e das avaliações desses modelos de detecção. Na prática, muitos *datasets* atuais são construídos a partir de *datasets* mais antigos [Sharma and Rattan, 2021], o que 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 e popular entre as fontes de dados de *malwares* Android [Singh and Khare, 2021]. O mesmo ocorre com diversos outros conjuntos de dados, como o Android Botnet, formado por dados do Malware Genome Project e VirusTotal. Além disso, há *datasets*, como o CICInvesAndMal2019, que afirmam incluir dados recentes, de 2019, porém contêm apenas características de versões bastante antigas (e.g., 2016 e inferior) da API do Android.

Dado este cenário de *datasets* utilizadas para detecção de *malware* 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. Além disso, diferente de estudos similares, como o [Kouliaridis et al., 2020], onde os autores limitam-se ao estudo de 10 *datasets*, 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*, 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 nos Anexos A e B.

As fontes de dados avaliadas 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.

Assim, 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 apenas 10 *datasets*. Nossa classificação de disponibilidade cobre todos os 84 *datasets* analisados, como pode ser visto na Tabela 1 e no Anexo A.

Dentre os repositórios de acesso restrito, o AndroZoo requer uma solicitação de chave de acesso, feita via email para o time do projeto. Outros, como o CIC-AndMal2017, exigem o preenchimento de um formulário contendo algumas perguntas sobre a utilização e identificação do *dataset*. 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.

Para a classificação das fontes, foram realizados testes de acesso em cada link, verificando sua validade e definindo se a fonte é uma loja, repositório, *dataset* ou até mesmo repositório e *dataset* simultaneamente. Dada essa classificação, apresentamos um panorama das informações sobre os mercados de aplicativos, *datasets* e repositórios de APKs a seguir.

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 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.

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

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, existem 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, assim, foram classificadas como indisponíveis. Como exemplos de outros casos de indisponibilidade, podemos citar a loja Hiapk, onde não foi encontrada qualquer forma de realizar o *download* dos APKs através do site, e a loja Eoemarket, onde os links disponibilizados para *download* retornam erro de página não encontrada.

A relação completa das lojas, incluindo *links* de acesso e informação sobre disponibilidade, podem ser encontradas na Tabela do Anexo A.

2.2. *Datasets* e Repositórios de APKs

A classificação quanto a atualização dos *datasets* e repositórios de APKs foi realizada utilizando três 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 no Anexo B.

Dos 45 *datasets* e repositórios, 25 são disponíveis, 11 são restritos e 9 são indisponíveis. Desse 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 da 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 2. 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.

A classificação de uma fonte de dados quanto ao acesso (disponível, restrito ou indisponível) define a dificuldade de se obter seu conteúdo. Isso é um fator que pode impactar no processo de validação e reprodução de uma solução baseada em aprendizado de máquina, pois o *dataset* é a peça central no processo. Empiricamente e de acordo com [Sharma and Rattan, 2021], observamos que os *datasets* mais utilizados em trabalhos na literatura são aqueles disponíveis publicamente ou de modo restrito, mas sem custos para os usuários.

Fontes restritas podem demorar mais para serem acessadas do que as disponíveis, pois, nesses casos, o acesso depende de alguma categoria de permissão. Já as fontes indisponíveis impedem a reprodutibilidade de trabalhos que as usaram. Entretanto, em-

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.

piricamente, observamos uma tendência crescente no sentido de priorizar a utilização de fontes publicamente disponíveis, como é o caso do Contágio Mobile e VirusShare [Sharma and Rattan, 2021].

Há alguns casos de fontes de dados disponíveis que disponibilizam tanto *datasets* quanto repositórios de APKs, caso da CICMalDroid 2020. Outra característica dessa fonte é ser 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.

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 no Anexo A.

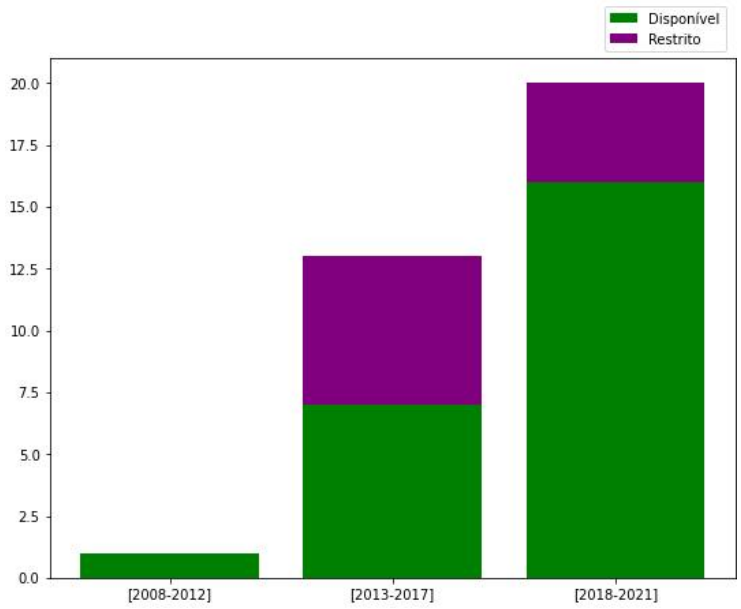

Figura 1. Atualização de Fontes de Dados Disponíveis ou Restritas

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, fundamentais para a reprodutibilidade dos trabalhos.

As próprias lojas acabam sendo também um obstáculo para a reprodutibilidade dos trabalhos. Grande parte delas impede a consulta a dados de versões antigas dos aplicativos, pois não mantêm ou disponibilizam as versões históricas dos aplicativos. Esse é caso de lojas como a Google Play Store, utilizada por 17 dos 35 trabalhos analisados. Algumas exceções, como as lojas como FDroid e FreewareLovers, permitem o download de versões mais antigas dos aplicativos. Portanto, para fins de reprodutibilidade, podemos concluir que o ideal seria os autores disponibilizarem um repositório com os APKs utilizados na pesquisa.

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 2. 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.

Utilizar dados antigos pode ser considerado um problema grave, pois pode

comprometer a qualidade e o desempenho dos modelos de aprendizado de máquina utilizados na detecção de *malwares*. De fato, como mostramos em outro trabalho [Vilanova et al., 2021] , *datasets* e dados antigos, quando utilizados no treinamento de modelos, possuem um impacto sobre o resultado dos modelos quando colocados em execução em um cenário atual, isto é, apenas com dados atuais.

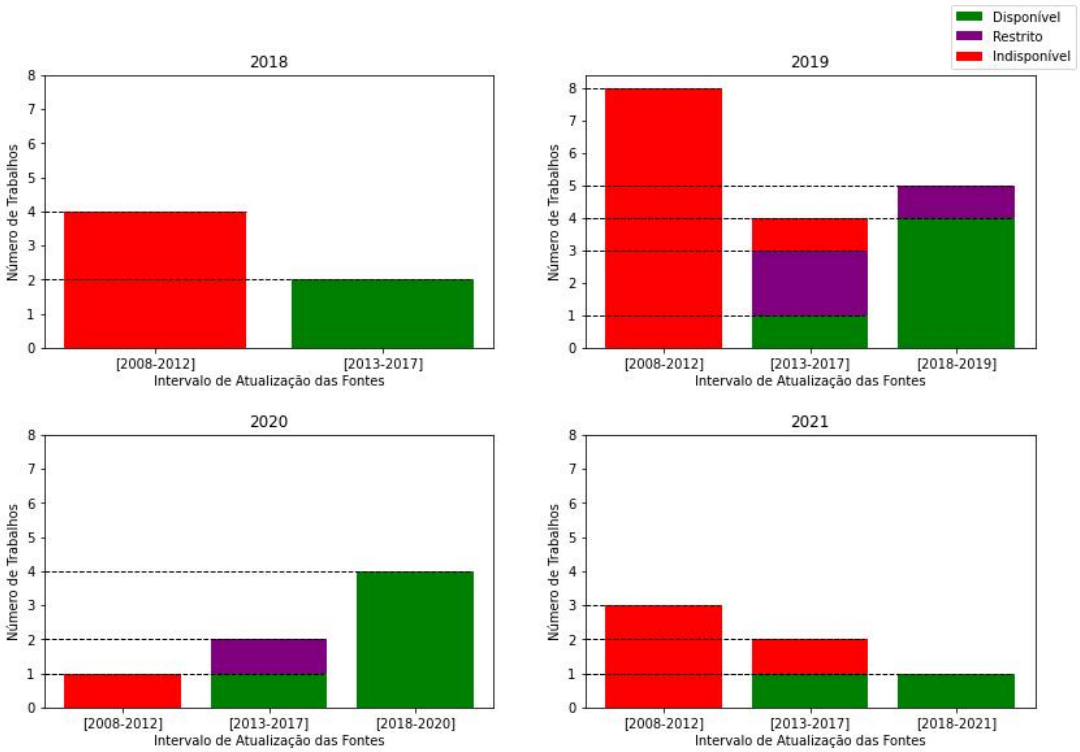

Figura 2. Atualização e Disponibilidade das Fontes de Dados Utilizadas

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]. Ambas 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

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

aplicações de APIs atuais, mas a 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.

A não coincidência das datas acarreta que as classificações quanto a atualidade das fontes de dados analisadas nesse trabalho sejam questionáveis, pois, para termos certeza do quão atual essas fontes são, seria necessária uma análise das versões das APIs presentes nelas. A diferença das datas de atualização de uma fonte de dados e das APIs que compõem tal fonte sugere que a data que deve ser levada em consideração para definir o período a qual pertence determinada fonte de dados seja definida pelas APIs das aplicações que a compõem.

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. Além disso, é importante destacar que a reprodutibilidade dos trabalhos é comprometida pelo uso de fontes indisponíveis e pela falta de detalhamento do *dataset* utilizado.

Um ponto importante percebido durante a realização desse trabalho é a possível inconsistência nas datas de atualização das fontes de dados. Essa inconsistência é um bom marcador para ser analisado em trabalhos futuros.

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 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.

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.

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.

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

Ming, F., Ting, L., Jun, L., Xiapu, L., Le, Y., and Xiaohong, G. (2020). Android malware detection: A survey. *Scientia Sinica Informationis*, 50(8):1148–1177.

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

Singh, G. and Khare, N. (2021). A survey of intrusion detection from the perspective of intrusion datasets and machine learning techniques. *International Journal of Computers and Applications*, 0(0):1–11.

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. In *VI Workshop Regional de Segurança da Informação e de Sistemas Computacionais (WRSeg)*, Charqueadas-RS, Brasil.

Soares, T., Siqueira, G., Barcellos, L., Sayyed, R., Vargas, L., Rodrigues, G., Assolin, J., Pontes, J., Feitosa, E., and Kreutz, D. (2021b). Detecção de Malwares Android: datasets e reprodutibilidade. In *VI Workshop Regional de Segurança da Informação e de Sistemas Computacionais (WRSeg)*, Charqueadas-RS, Brasil.

SophosLabs (2021). Sophos 2021 threat report. https://www.sophos.com/en-us/medialibrary/pdfs/technical-papers/sophos-2021-threat-report.pdf.

Vilanova, L., Sayyed, R., Soares, T., Siqueira, G., Rodrigues, G., Feitosa, E., and Kreutz, D. (2021). Análise do impacto de viés nos conjuntos de dados para detecção de malwares android. In *VI Workshop Regional de Segurança da Informação e de Sistemas Computacionais (WRSeg)*, Charqueadas-RS, Brasil.

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.

Wei, F., Li, Y., Roy, S., Ou, X., and Zhou, W. (2017). Deep ground truth analysis of current android malware. *Detection of Intrusions and Malware, and Vulnerability Assessment*, 10327.

Yan, P. and Yan, Z. (2018). A survey on dynamic mobile malware detection. *Software Quality Journal*, 26(3):891–919.

A. Anexo 1

A Tabela 2 apresenta os links e a informação de disponibilidade dos 39 mercados de aplicativos analisados.

Tabela 2. Lojas de Aplicativos

Lojas	***Link***	**Disponibilidade**
SlideME	https://slideme.org/	Disponível
Apkmirror	https://www.apkmirror.com/	
AppsApk	https://www.appsapk.com/	
ChinaMobile	https://www.bityli.com/nNzQ3	
Coolapk	https://coolapk.com/	
FDroid	https://f-droid.org/	
Flyme	https://app.flyme.cn/	
Appchina	https://www.appchina.com/	
Mumayi	http://www.mumayi.com	
Anzhi Store	http://www.anzhi.com	
FreewareLovers	https://www.freewarelovers.com/	
ProAndroid	https://proandroid.net/	
1mobile	https://1mobile.market/	
AndroidLista	https://www.androidlista.com/	
Tencent App Market (T-Market)	https://sj.qq.com/myapp/	
Wandoujia	https://www.wandoujia.com/apps	
Lenovo MM	https://www.lenovomm.com/	
Sogou	https://zhushou.sogou.com/	
360	https://zhushou.360.cn/	
Baidu	https://shouji.baidu.com/	
Huawei	https://appgallery.huawei.com	
Mi	https://app.mi.com/	
163 app store	https://m.163.com/	
YingYongBao Store	https://android.myapp.com/	
91 app store	https://play.91.com/	
2mm AndroidDrawer	https://www.androiddrawer.com/	Indisponível
Eoemarket	www.eoemarket.com/	
GetJar	www.getjar.com/	
GFan	https://apk.gfan.com/	
Wangyi	https://m.163.com/apps	
PandaApp	https://android.pandaapp.com	
Hiapk	www.hiapk.com/	
AnGeeks	www.angeeks.com/	
Anruan	www.anruan.com/	
AndroidBest	https://androidbest.ru/	
AndroidLife	https://androidlife.ru/	
App Gionee	https://www.appgionee.com/	
10086	http://www.10086.cn/index_5074.htm	
NQ Mobile/Link Motion	http://en.lkmotion.com/	

B. Anexo 2

A Tabela 3 traz, além dos links de acesso na coluna *Link*, o ano da última atualização de cada fonte de dados na coluna *Última Atualização* e onde essa informação foi encontrada (i.e., site ou *paper*) em *Origem da data*. Fontes para quais as informações não foram encontradas através de nossas buscas, receberam "-" na tabela.

Tabela 3. *Datasets* e Repositórios de APKs

***Datasets* e Repositórios de APKs**	**Última Atualização**	**Origem da data**	***Link***
Ether Malware Analysis Dataset	2008	Site	www.azsecure-data.org/other-data.html
Contagio Malware Dump	2013	Site	https://bityli.com/19nPy
CIC-AAGM2017	2017	Site	https://bit.ly/39d9npX
MudFlow	2016	Site	https://bit.ly/3k9GrWf
Android Botnet	2015	Site	https://bit.ly/3hQcjOn
M0Droid	2014	Site	www.azsecure-data.org/other-data.html
GaziBenignApp	2017	Site	https://bit.ly/399Ug0J
Contagio Mobile	2018	Site	https://contagiominidump.blogspot.com
Android Permissions Dataset	2018	Site	https://data.mendeley.com/datasets/b4mxg7ydb7/1
VirusShare	2018	Site	https://bit.ly/2YXPgug
CICInvesAndMal2019	2019	Site	https://bit.ly/2VIQk3W
TheZoo	2015	Site	https://thezoo.morirt.com
AndroMalShare	2021	Site	https://malshare.com/index.php
Koodus	2021	Site	https://docs.koodous.com/
Drebin-215	2018	Site	https://bit.ly/2YXOu0j
Dataset of Android Permissions	2018	Site	https://bit.ly/3zdp1fL
CICMalDroid2020	2020	Site	https://bit.ly/3lpYgzW
Android Malware and Normal Permissions Dataset	2018	Site	https://data.mendeley.com/datasets/958wvr38gy/5
Android Malware and Benign Application Dataset	2020	Site	https://data.mendeley.com/datasets/b4mxg7ydb7/3
PARUDroid	2020	Site	https://data.mendeley.com/datasets/mg5c8jxbhm/2
Comodo Cloud Security Center	2021	Site	https://bit.ly/2YXePf1
Wang's Repository	2018	Site	https://infosec.bjtu.edu.cn/wangwei/?page_id=85
Drebin4000	2021	Site	https://bit.ly/3zalrD9
AMD6000	2021	Site	https://bit.ly/3zalrD9
CICAndMal2017	2017	Site	https://www.unb.ca/cic/datasets/andmal2017.html
CCCS-CIC-AndMal-2020	2020	Site	www.unb.ca/cic/datasets/andmal2020.html
AndroZoo	2016	Site	https://androzoo.uni.lu/access
Andro-Dumpsys	2016	Site	https://ocslab.hksecurity.net/andro-dumpsys
Andro-AutoPsy	2015	Site	https://ocslab.hksecurity.net/andro-autopsy
Andro-Profiler	2016	Site	https://ocslab.hksecurity.net/andro-profiler
Andro-Tracker	2015	Site	https://ocslab.hksecurity.net/andro-tracker
UpDroid	2018	Paper	https://bit.ly/3hw3ZmH
Contagio Mini Dump	2018	Site	http://contagiominidump.blogspot.com
COVID19 Apps	2021	Paper	https://zenodo.org/record/4660140#.YTUWpxvPzIU
Heldroid	2015	Paper	https://github.com/necst/heldroid
DroidKin	2015	Paper	-
MobileSandbox project	2016	Site	https://bit.ly/398Wmhb
The Drebin Dataset	2012	Site	www.sec.cs.tu-bs.de/$\sim$danarp/drebin/
Android Malware Genome Dataset	2012	Site	https://bit.ly/3EkiXWr
Android PRAGuard Dataset	2015	Paper	https://bit.ly/3lsNbhD
PlayDrone Project	2014	Paper	https://archive.org/details/android_apps&tab=about
McAfee	-	-	-
Inter-Component Communication Repository (IcRE)	-	-	-
New Malware Families 2015	-	-	-
VirusTotal Malware Service Intelligence	2021	Site	https://bit.ly/3k8caqL
Kharon Malware Dataset	2019	Site	http://kharon.gforge.inria.fr/dataset/