Compilação de um Dataset especializado em treinamento de uma ML para detecção de malwares em sistemas Android

Jessica de Figueredo Colares, Lucca Dourado Cunha

[1]Instituto de Computação – Universidade Federal do Amazonas (UFAM)
Av. Gen. Rodrigo Octávio, 6200, Coroado I, Setor Norte
Campus Universitário – 69080-900 – Manaus – AM

jessica.colares@icomp.ufam.edu.br, lucca.dourado@icomp.ufam.edu.br

1. Introdução

A segurança cibernética é um tema que tem ganhado uma importância cada vez maior na era digital. Com a crescente ubiquidade dos dispositivos *Android* na vida cotidiana, o seu papel tornou-se fundamental. A base de usuários de dispositivos *Android* continua a expandir-se, e, juntamente com ela, a proliferação de aplicativos móveis. No entanto, à medida que esse ecossistema se expande, também aumenta exponencialmente o risco de infecção por *malware*. Neste contexto, a detecção de *malware* torna-se uma tarefa essencial para proteger a privacidade e a integridade dos dados dos usuários.

A complexa tarefa de detecção de malware envolve análise detalhada de código e comportamento. A aplicação de técnicas de aprendizado de máquina automatiza parte desse processo, capacitando modelos a reconhecer padrões suspeitos e alertar os usuários. No entanto, a eficácia dessas ferramentas depende de conjuntos de dados especializados e confiáveis, uma lacuna evidenciada na pesquisa [Mahindru and Sangal 2021].

Este trabalho propõe uma ferramenta atualizada de detecção de malware em dispositivos Android, utilizando avançadas técnicas de machine learning. A criação de um repositório de malwares atualizado é fundamental para o treinamento de modelos capazes de identificar até as variantes mais sutis de malware.

O treinamento com base nesse conjunto de dados será comparado a conjuntos mais comuns, avaliando eficiência, acurácia e incidência de falsos positivos [Akhtar 2023]. Essa análise aprofundada guiará melhorias para futuros desenvolvimentos, visando aumentar a utilidade e confiabilidade dos conjuntos de dados.

Além de contribuir para a segurança dos dispositivos Android, este trabalho impacta a criação de um ambiente digital mais seguro para todos os usuários. Ao fortalecer a segurança cibernética, fortalecemos a confiança dos usuários na utilização de seus dispositivos Android.

A relevância global desta pesquisa destaca-se no contexto da segurança cibernética em um mundo interconectado. O desenvolvimento de técnicas avançadas de detecção de malware beneficia não apenas indivíduos, mas também empresas, governos e organizações em todo o mundo. Este trabalho é uma contribuição significativa para a mitigação de riscos cibernéticos em um ambiente cada vez mais interligado. Os participantes deste trabalho estão comprometidos em tornar o mundo digital um lugar mais seguro.

2. Revisão da literatura

O *Android* vem sendo bastante estudado principalmente no âmbito de segurança. Depois da revolução que os *smartphones* vieram causando no nosso mundo, grande parte da população confia neles informações e funções de extrema sensibilidade como aplicativos bancários [Cho et al. 2013], o estudo e a exposição dessas vulnerabilidades é importante para definir quais medidas tomar e o que implementar para resguardar a integridade do sistema [Amaral et al. 2016]. Uma falha comum de segurança que ocorre com o sistema é a concessão de permissões excessivas a aplicativos, tornando-os possíveis ameaças a integridade do sistema, esse problema é originado de desenvolvedores dos quais não conseguem categorizar com qualidade quais permissões são essenciais para o funcionamento do mesmo [Jacobsen 2015]. Muitas das vulnerabilidades acabam sendo criadas pelo próprio usuário final por um desconhecimento do funcionamento do sistema e falta de praticidade com as soluções existentes [Rotondo 2016]. A desatualização do sistema também é fator extremamente limitante causado pelo usuário ou até uma limitação de hardware do dispositivo [NASCIMENTO and Cintra 2019]. Há uma falta de aplicativos intuitivos e centralizados em segurança para usuários não familiarizados com tecnologia. Portanto, uma aplicação desse tipo seria benéfica ao simplificar o gerenciamento de questões de segurança e vulnerabilidades do sistema.

O entendimento dos atuais sistemas é essencial para direcionarmos a evolução nos quesitos que as fraquezas são aparentes e onde há um bom campo para evolução [Braga et al. 2012]]. A forma como a rede sem fio evoluiu com o passar do tempo é uma grande prova deste fato, atualmente temos inúmeras técnicas já aplicadas no sistema para proteção de códigos maliciosos que o sistema originalmente não teriam defesas sobre [Scota et al. 2018] [de Oliveira and Carro 2014]. O estudo comparativo também se mostra extremamente eficiente na evolução de todos os envolvidos [Oliveira Costa and Duarte Filho 2013], comparando o nosso sistema de estudo com os concorrentes de mercado podemos levantar suas vantagens e desvantagens e analisar a aplicação de boas práticas de um sistema a outro. Tendo em mente até onde fomos nas atuais soluções e com o estudo de soluções, é perceptível a direção em que a comunidade está tendendo para o futuro: soluções de detecção de *malwares* por meio de *machine learning*. Essa recente área consoante as pesquisas mais recentes tem um grande potencial prático para todos os usuários e é uma área com o caminho a frente bem definido [Liu et al. 2020], portanto é um assunto bastante pertinente a ser tratado para a nossa solução.

A relevância dos estudos focados diretamente nos *malwares* e suas técnicas de exploração de fraquezas do sistema se encontram equiparadas às pesquisas ligadas a sistemas de segurança diretamente [Inácio and Gomes 2014]. Estudos direcionados a categorizar e apresentar padrões nos *malwares* desenvolvidos nos últimos anos nos mostram uma que sempre há uma tendência na forma como os ataques ocorrem conforme o momento [Zhou and Jiang 2012], é possível fazer uma correlação do que ocorreu no passado com o que ocorre hoje Um exemplo da efetividade deste estilo de estudo é que com eles surgiram aplicações com a intenção de identificar *malwares*, eles se utilizam de técnicas como regras de associação e qualidade de regras [da Silva Rocha et al. 2022], análise de processos [Rotondo 2016] e construção de *datasets* [Vilanova et al. 2022] em conjunto com essa categorização dos *malwares* para aumentar sua eficácia.

3. Andamento dos Vetores

Devido ao tempo limitado de pesquisa, concentramos esforços na criação de uma versão inicial do conjunto de dados para treinamento, com o objetivo de adquirir uma compreensão mais aprofundada dos elementos essenciais para o treinamento de uma máquina virtual. Buscamos identificar os requisitos mínimos necessários para um treinamento eficaz de um detector de malwares e compilamos um conjunto de dados composto por uma extensa lista de permissões. Essa lista serve para determinar se um aplicativo analisado é benigno ou malicioso, possibilitando que a inteligência artificial identifique padrões nas permissões e faça previsões sobre possíveis aplicativos que possam explorar esses acessos. Todas as informações adicionais foram removidas do banco de dados gerado, priorizando a simplicidade e eficácia na análise.

Prosseguindo com esta pesquisa, nossa intenção é criar diversas versões do dataset, incorporando dados mais específicos em cada uma delas. O objetivo é realizar testes individuais dessas versões em várias inteligências artificiais, identificando quais dados contribuem para a análise eficaz do detector e quais podem ser considerados dispensáveis. Após essa etapa, almejamos aprimorar e otimizar o dataset, reduzindo ruídos, para, por fim, comparar se os benefícios de um conjunto de dados mais refinado e elaborado compensam em relação a um dataset convencional e pouco refinado. Este processo permitirá avaliar se os ganhos na qualidade do dataset são significativos.

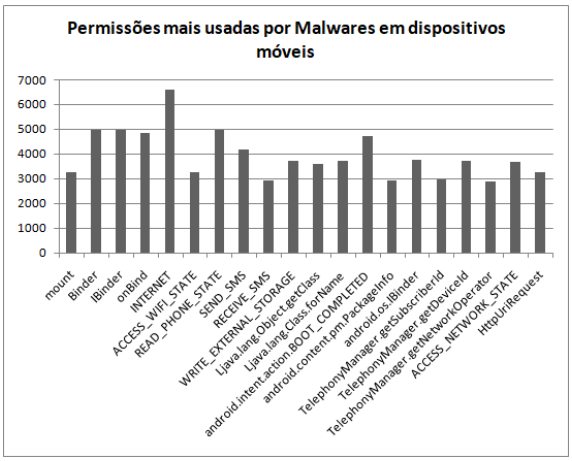

Figura 1. Gráfico de colunas com o objetivo de informar as 20 permissões mais utilizadas pelos malwares em dispositivos móveis.

Com base no dataset criado, procedeu-se à classificação de cada um dos aplicativos analisados. Os aplicativos benignos foram designados como *B*, enquanto os aplicativos maliciosos, ou malwares, foram categorizados como *S*. Foram excluídos da lista todos os aplicativos classificados como benignos, resultando apenas na análise dos malwares.

Em seguida, foi realizado um levantamento das permissões utilizadas por cada malware no conjunto de dados. No dataset, o valor *0* indica que a permissão não foi utilizada, enquanto o valor *1* indica o uso da permissão. Foi realizada uma contagem para cada permissão, revelando quantos malwares fizeram uso dela.

Na Tabela 1, apresentamos as 20 permissões mais frequentemente utilizadas por malwares em sistemas Android. Foram minuciosamente analisadas 215 permissões, destacando a INTERNET como a preferida dos malwares, com 6572 instâncias fazendo uso

Tabela 1. Número de permissões mais utilizadas pelos Malwares em dispositivos móveis

Permissões	Malwares
mount	3265
Binder	4959
IBinder	4947
onBind	4815
INTERNET	6572
ACCESS_WIFI_STATE	3236
READ_PHONE_STATE	4951
SEND_SMS	4178
RECEIVE_SMS	2929
WRITE_EXTERNAL_STORAGE	3727
Ljava.lang.Object.getClass	3584
Ljava.lang.Class.forName	3702
android.intent.action.BOOT_COMPLETED	4691
android.content.pm.PackageInfo	2931
android.os.IBinder	3758
TelephonyManager.getSubscriberId	2966
TelephonyManager.getDeviceId	3688
TelephonyManager.getNetworkOperator	2862
ACCESS_NETWORK_STATE	3684
HttpUriRequest	3261

desta autorização. Esses resultados apontam para a vulnerabilidade significativa da camada de internet nos dispositivos móveis.

4. Conclusão

A constatação de que a internet é a parte mais suscetível dos dispositivos móveis ressalta a urgência de aprimorar as medidas de segurança no acesso a essa rede, especialmente para proteger dados sensíveis dos usuários do sistema Android. Diante desse cenário, uma abordagem proativa seria implementar melhorias no sistema de detecção de uso mal-intencionado da permissão de acesso à internet.

Este trabalho, ao criar um dataset abrangente, fornece uma base sólida para a implementação de Inteligência Artificial no sistema do dispositivo. Tal integração possibilitaria uma detecção mais otimizada de comportamentos maliciosos relacionados à permissão de acesso à internet. Esse avanço na segurança não apenas protegeria os usuários contra ameaças cibernéticas, mas também contribuiria para a criação de um ambiente digital mais seguro e confiável para os dispositivos Android.

Portanto, a utilização sistemática desses estudos direcionados a *malwares* nos mostra que podem alavancar a eficiência de qualquer solução, e que a compreensão das atuais tendências dos desenvolvedores mal-intencionados é de suma importância para possibilitar a contenção de possíveis brechas de segurança.

Referências

Akhtar, M. S. (2023). Analyzing and comparing the effectiveness of various machine learning algorithms for android malware detection. *Advances in Mobile Learning Educational Research*, 3(1):570–578.

Amaral, G., Silva, R., Rotondo, G., and Amaral, É. (2016). Um estudo sobre vulnerabilidades do android: Ferramentas e soluções para o usuário. *Anais SULCOMP*, 8.

Braga, A. M., do Nascimento, E. N., da Palma, L. R., and Rosa, R. P. (2012). Introdução à segurança de dispositivos móveis modernos–um estudo de caso em android. *Sociedade Brasileira de Computação*.

Cho, T., Kim, Y., Han, S., and Seo, S.-H. (2013). Potential vulnerability analysis of mobile banking applications. In *2013 International Conference on ICT Convergence (ICTC)*, pages 1114–1115. IEEE.

da Silva Rocha, V., Kreutz, D., Pontes, J., and Feitosa, E. (2022). Avaliaçao de métodos de classificaçao baseados em regras de associaçao para detecçao de malwares android. In *Anais do XXII Simpósio Brasileiro em Segurança da Informação e de Sistemas Computacionais*, pages 316–329. SBC.

de Oliveira, B. M. and Carro, S. A. (2014). Desenvolvimento de um framework para o monitoramento de dispositivos móveis na plataforma android. In *Colloquium Exactarum. ISSN: 2178-8332*, volume 6, pages 80–98.

Inácio, L. R. and Gomes, A. R. L. (2014). Uma abordagem sobre problemas de segurança da informação por meio do desenvolvimento de aplicações maliciosas para android. *e-xacta*, 7(2):87–106.

Jacobsen, W. (2015). Uma solução para avaliação de riscos em aplicativos para dispositivos móveis.

Liu, K., Xu, S., Xu, G., Zhang, M., Sun, D., and Liu, H. (2020). A review of android malware detection approaches based on machine learning. *IEEE Access*, 8:124579–124607.

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

NASCIMENTO, F. R. d. and Cintra, F. G. (2019). Vulnerabilidades de segurança em dispositivos android: análises e estatísticas (2009-2019).

Oliveira Costa, N. P. and Duarte Filho, N. F. (2013). Análise e avaliação funcional de sistemas operacionais móveis: Vantagens e desvantagens. *Revista de Sistemas e Computação-RSC*, 3(1).

Rotondo, G. (2016). Prosec-uma solução para o controle de processos maliciosos na plataforma android.

Scota, D. F., de Andrade, G. E., and da Costa Xavier, R. (2018). Configuraçao de rede sem fio e segurança no sistema operacional android.

Vilanova, L., Kreutz, D., Assolin, J., Quincozes, V., Miers, C., Mansilha, R., and Feitosa, E. (2022). Adbuilder: uma ferramenta de construçao de datasets para detecçao de malwares android. In *Anais Estendidos do XXII Simpósio Brasileiro em Segurança da Informação e de Sistemas Computacionais*, pages 143–150. SBC.

Zhou, Y. and Jiang, X. (2012). Dissecting android malware: Characterization and evolution. In *2012 IEEE symposium on security and privacy*, pages 95–109. IEEE.