Uma GUI para Hackers do Bem Aprenderem Sobre Malwares Sintéticos

Leonardo Karling Sonco[1,3], Angelo Nogueira[1,2,3], Diego Kreutz[1,2,3], Rodrigo Brandão Mansilha[1,2,3]

[1]Laboratório de Estudos Avançados em Computação (LEA)

[2]Programa de Pós-Graduação em Engenharia de Software (PPGES)

[3]Universidade Federal do Pampa (UNIPAMPA)

{leonardosonco.aluno, angelonogueira.aluno, diegokreutz, mansilha}@unipampa.edu.br

Resumo. *Hackers do Bem precisam se manter atualizados para enfrentar as técnicas emergentes usadas por Hackers do Mal. Entre essas técnicas, está o uso de aprendizado profundo para criar malwares Android que imitam o comportamento de aplicativos legítimos, com o objetivo de enganar antivírus enquanto exploram novas vulnerabilidades. Nesse cenário, desenvolvemos o Malware DataLab, uma plataforma dedicada ao ensino de técnicas de aprendizado profundo, com o objetivo de ampliar datasets de malwares Android utilizando dados sintéticos. Este trabalho apresenta a interface gráfica do MalSynGen, a ferramenta de geração de dados tabulares sintéticos do Malware DataLab. Uma avaliação experimental preliminar demonstra o impacto positivo da proposta.*

Abstract. *Ethical hackers need to constantly stay updated to combat emerging techniques used by malicious hackers. This includes the use of deep learning to create new Android malware that mimics the behavior of legitimate applications, aiming to deceive antivirus systems while exploiting newly discovered vulnerabilities. In this context, we developed Malware DataLab, a platform focused on teaching deep learning techniques with the goal of expanding Android malware datasets using synthetic data. This work presents the graphical interface of MalSynGen, the tool for generating synthetic tabular data within the Malware DataLab. A preliminary experimental evaluation demonstrates the positive impact of the proposal.*

1. Introdução

Nos últimos anos, as Redes Neurais Artificiais (RNAs) têm se destacado como uma tecnologia relevante, com aplicações em diversas áreas do conhecimento humano [Fleck et al. 2016]. Um exemplo são as *Generative Adversarial Networks* (GANs), que permitem a geração de dados artificiais, proporcionando desde a melhoria na qualidade de imagens até o aumento no desempenho de tecnologias analíticas. No entanto, essa tecnologia também pode ser utilizada para a criação e disseminação de *malwares*, aumentando a complexidade da detecção por antivírus [Hu and Tan 2017].

O combate contra esses *malwares* dependem muito da qualidade e quantidade de dados de treinamento, que muitas vezes estão obsoletos pela dificuldade e demora

de gerar novos dados [Miranda et al. 2022]. Para resolver esse problema existem algumas técnicas que utilizam de *conditional* GANs (cGANs) [Nogueira et al. 2024a], como a DroidAugmentor [Casola et al. 2023] e, mais recentemente, a MalSynGen [Nogueira et al. 2024b]. Contudo, a execução em escala dessas ferramentas oferece um desafio significativo devido à complexidade computacional e aos requisitos de hardware. Essas barreiras podem ser superadas com ajuda de soluções como a AutoDroid [Laviola et al. 2023], um sistema que permite disponibilizar ferramentas como a MalSynGen através de virtualização leve. Dessa forma, uma série de complexidades subjacentes de configuração de ambiente se tornam transparentes para o usuário.

Uma limitação da MalSynGen (e da AutoDroid) é a ausência de uma interface gráfica para facilitar seu uso. A falta dessa interface dificulta a interação com usuários não técnicos, aumentando a complexidade de operação do sistema. Isso pode dificultar o aprendizado e elevar o risco de erros, o que pode desestimular usuários com menos experiência. Essa limitação é particularmente relevante e desafiadora no contexto de projetos como o Malware DataLab[1], que tem como objetivo desenvolver um ambiente de ensino e aprendizagem de redes neurais aplicadas à cibersegurança. O Malware DataLab é um projeto apoiado pela Rede Nacional de Ensino e Pesquisa (RNP)[2], no âmbito do Programa Hackers do Bem[3], que visa contribuir para a formação de profissionais na área de cibersegurança.

O objetivo deste trabalho é apresentar uma Interface Gráfica do Usuário (*Graphical User Interface* - GUI) para a MalSynGen, denominada GUI4MalSynGen. Ela oferece as funcionalidades básicas necessárias para que usuários sem experiência técnica, mas interessados em aprender sobre tecnologias e processos relacionados às RNAs, possam utilizá-la. Essa iniciativa visa tornar a MalSynGen mais acessível a usuários que não estão familiarizados com tecnologias de virtualização e interfaces de linha de comando, ampliando seu potencial de impacto. A GUI4MalSynGen foi avaliada utilizando o método TAM, e os resultados indicam o potencial da proposta. A GUI4MalSynGen está publicamente disponível[4], permitindo a execução autônoma (*stand-alone*) dos experimentos de execução da MalSynGen através da AutoDroid. Além disso, a GUI4MalSynGen serviu como base para o desenvolvimento de um serviço acessível na nuvem para alunos credenciados no Malware DataLab.

No restante deste trabalho, apresentamos a solução proposta na Seção 2, seguida de uma primeira avaliação na Seção 3, e encerramos com as considerações finais na Seção 4.

2. GUI4MalSynGen

2.1. Arquitetura

A Figura 1 apresenta uma visão de sistema da solução em notação C4[5]. Ela ilustra quatro componentes principais: o usuário (topo à esquerda), a proposta (topo à

[1] `https://malwaredatalab.github.io/`
[2] `https://rnp.br`
[3] `https://hackersdobem.org.br`
[4] `https://github.com/MalwareDataLab/GUI4MalSynGen`
[5] `https://c4model.com/`

direita), o AutoDroid componente a ser requisitado (base à esquerda) por fim a MalSynGen (base à direita).

O usuário interage com a ferramenta AutoDroid por meio da GUI4MalSynGen, uma interface que permite acompanhar o progresso e o desempenho do treinamento da RNA. A AutoDroid é responsável por receber o *dataset* e as configurações de parâmetros fornecidos pelo usuário, repassando-os ao MalSynGen para a geração dos dados.

A arquitetura da GUI4MalSynGen foi projetada com uma clara separação de responsabilidades, facilitando a manutenção do código. Além disso, a interface proporciona uma interação clara e funcional, permitindo que o usuário execute suas tarefas de maneira intuitiva e eficaz.

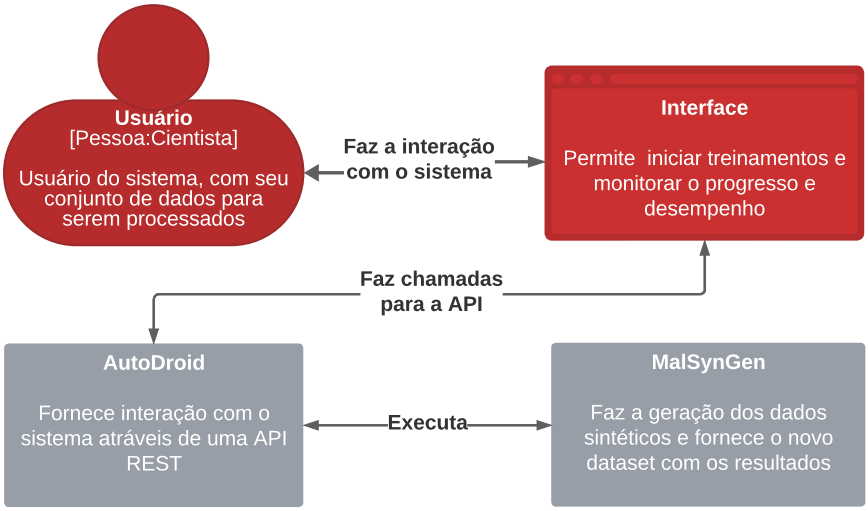


Figure 1. Diagrama de contexto da GUI4MalSynGen.

2.2. Descrição das Telas

A GUI4MalSynGen é composta por quatro telas principais: Tela Inicial, Tela "Entenda as Ferramentas", Tela de Execução e Tela de Resultados. Cada uma foi projetada para oferecer uma experiência de usuário fluida e intuitiva, atendendo aos requisitos funcionais da aplicação. Além disso, as cores foram cuidadosamente selecionadas para melhorar a legibilidade e acessibilidade.

- **Tela Inicial**. Ao acessar a Tela Inicial, o usuário é recebido por uma interface minimalista com duas opções: "Entenda as Ferramentas" e "Ambiente de Execução". Essa organização facilita a navegação para usuários com diferentes níveis de conhecimento. O usuário pode optar por realizar um treinamento introdutório sobre AutoDroid e MalSynGen ou explorar diretamente as funcionalidades da ferramenta.
- **Tela "Entenda as Ferramentas"**. Esta tela oferece uma introdução aos termos essenciais para a compreensão do processo de geração de dados sintéticos. São apresentadas explicações sobre conceitos como Redes Neurais e

CGANs, além de uma breve descrição das ferramentas MalSynGen e AutoDroid. A ideia é que o usuário possa consultar esta tela antes de iniciar o ambiente de execução, proporcionando um primeiro contato com os conceitos e ferramentas para uma melhor compreensão e interação.

- **Tela Ambiente de Execução**. O desafio para projetar esta tela é organizar uma ampla gama de parâmetros, tornando a ferramenta acessível a usuários inexperientes, sem comprometer a usabilidade para usuários avançados. A tela foi dividida em três seções principais para facilitar a interação com as funcionalidades. Na primeira seção, "Parâmetros de Treinamento", o usuário pode selecionar uma campanha predefinida ou personalizar uma nova, além de escolher o conjunto de dados que será utilizado para gerar os dados sintéticos. A segunda seção exibe os parâmetros e conjuntos de dados selecionados para cada campanha. Na terceira seção, o usuário pode visualizar os processos em execução, com seus respectivos parâmetros e status. Ao selecionar um processo, o usuário é redirecionado para a Tela de Resultados.
- **Tela de Resultados**. Nesta tela são exibidos os resultados do processo selecionado. O principal desafio foi organizar os diversos resultados gerados pela MalSynGen de maneira acessível para usuários inexperientes. A tela apresenta gráficos principais que permitem avaliar o sucesso do treinamento, juntamente com as configurações utilizadas para referência. Os resultados são apresentados em ordem de importância e acompanhados de descrições para facilitar a interpretação, especialmente para iniciantes. Uma funcionalidade importante é a opção de exportar os dados artificiais gerados diretamente para o dispositivo do usuário, oferecendo controle total sobre os dados produzidos.

2.3. Implementação

A GUI4MalSynGen foi desenvolvida utilizando uma combinação de tecnologias para proporcionar flexibilidade e consistência. O React[6] foi escolhido para a construção da interface, pois sua capacidade de reutilização de componentes facilita a manutenção. Para a estilização, foi utilizado o Tailwind CSS[7], uma biblioteca que agiliza o processo de estilização dos componentes e aumenta a produtividade no desenvolvimento. Além disso, componentes prontos das bibliotecas MUI Material[8] e Swiper[9] foram integrados, otimizando ainda mais o desenvolvimento da interface.

Na comunicação com o AutoDroid, foi utilizado o TypeScript[10], uma linguagem de programação que adiciona tipos estáticos ao JavaScript, resultando em um código mais robusto e de fácil refatoração. Para realizar as requisições, foi empregada a biblioteca Axios[11], escolhida por sua API simplificada e sua compatibilidade com o uso de *async/await*, o que facilita a gestão de promessas e melhora a legibilidade do código.

[6] `https://react.dev/`
[7] `https://tailwindcss.com/`
[8] `https://mui.com/`
[9] `https://swiperjs.com/`
[10] `https://www.typescriptlang.org/`
[11] `https://axios-http.com/ptbr/docs/intro`

A Figura 2 apresenta exemplos de telas da GUI4MalSynGen. O código fonte e uma versão executável podem ser encontrados em repositório online público[4]. Na Figura 2(a), destacamos o *design* minimalista. Na Figura 2(b), apresentamos o conteúdo introdutório. Na Figura 2(c), o desafio é exibir parâmetros complexos de maneira acessível. Para isso, criamos recursos que permitem a seleção dos parâmetros mais relevantes, além de oferecer uma série de valores pré-configurados para facilitar o processo. Na Figura 2(d), ressaltamos a seleção dos resultados mais relevantes, a organização dos resultados em uma sequência lógica e a inclusão de dicas para auxiliar na interpretação.

3. Avaliação

Nesta seção, apresentamos a metodologia e os resultados de uma primeira avaliação da GUI4MalSynGen.

3.1. Metodologia

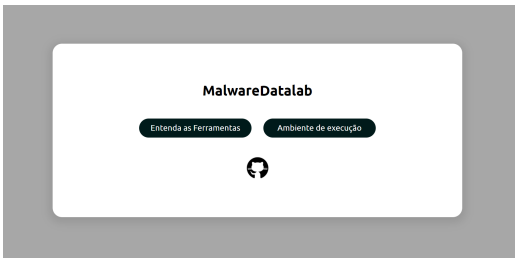
O questionário inclui um campo aberto e opcional para que os participantes possam adicionar comentários e sugestões, além de uma série de sentenças baseadas no *Technology Acceptance Model* (TAM) [Davis et al. 1989]. O método TAM propõe a formulação de sentenças objetivas, que são apresentadas aos usuários acompanhadas de uma escala Likert [Likert 1932]. Seguindo esse modelo, foram criados dois grupos de sentenças, cada um contendo cinco itens, conforme listado na Tabela 1. Para cada sentença, foi adotada uma escala Likert de 5 pontos, variando de "Discordo totalmente" a "Concordo totalmente".

Table 1. Grupos e sentenças do TAM aplicado.

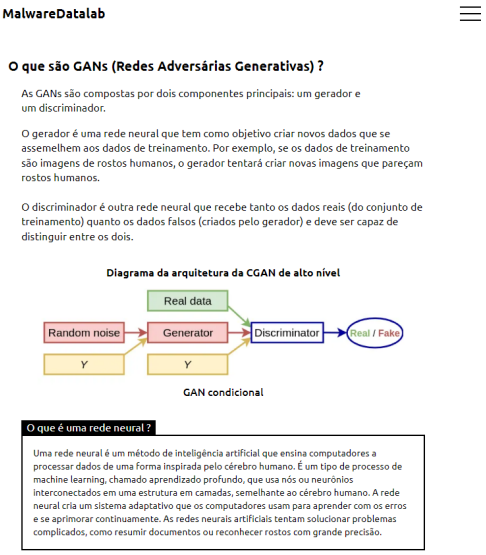
Grupo	#	Sentença
Utilidade Percebida (UP). Nível em que uma pessoa acredita que o uso do GUI4MalSynGen contribui para melhorar o seu desempenho no trabalho.	**UP1**	Usar a GUI4MalSynGen melhorou meu entendimento sobre cGANs.
	UP2	Usar a GUI4MalSynGen melhorou meu entendimento sobre o MalSynGen.
	UP3	Usar a GUI4MalSynGen economiza tempo.
	UP4	Usar a GUI4MalSynGen melhorou meu desempenho para gerar dados sintéticos.
	UP5	Eu considero útil a GUI4MalSynGen para a utilização do MalSynGen.
Facilidade de Uso (FU). Nível em que uma pessoa acredita que usar o GUI4MalSynGen é livre de esforço.	**FU1**	É fácil aprender a usar a GUI4MalSynGen.
	FU2	Minha interação com a GUI4MalSynGen é clara e compreensível.
	FU3	É fácil fazer o que eu quero com a GUI4MalSynGen.
	FU4	A GUI4MalSynGen se adapta facilmente às minhas necessidades de interação.
	FU5	O uso da GUI4MalSynGen requer pouco esforço mental.

O questionário foi disponibilizado online através da ferramenta Google Forms[12]. Na primeira etapa, os participantes são apresentados a um Termo de Consentimento

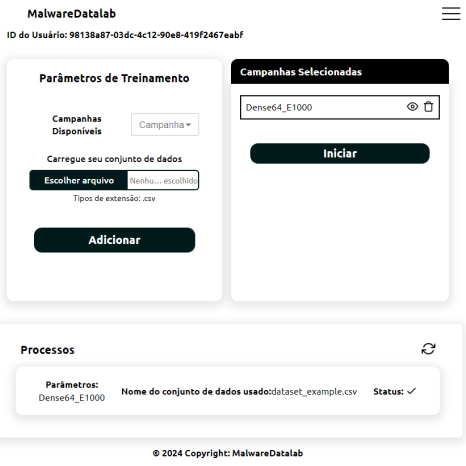
[12] `https://www.google.com/intl/pt-BR/forms/about/`



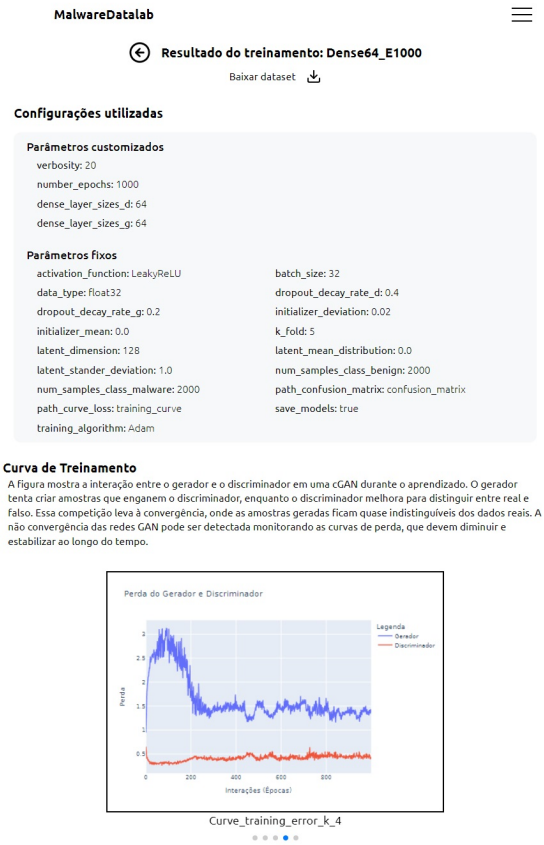
(a) Tela Inicial.



(b) Entenda as Ferramentas.



(c) Ambiente de Execução.



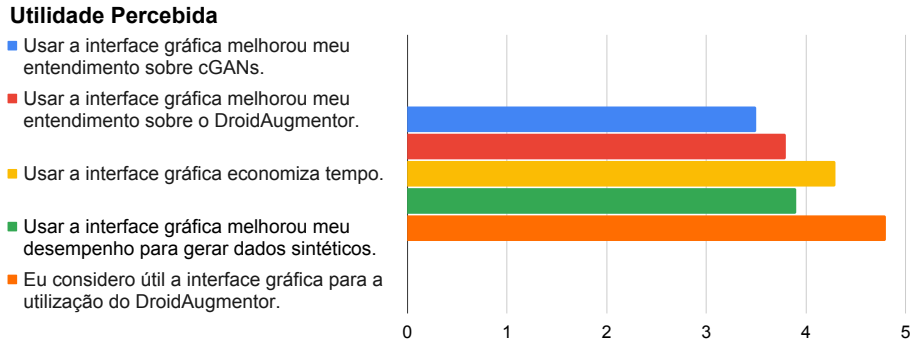
(d) Tela de Resultados.

Figure 2. Exemplos das telas principais da GUI4MalSynGen.

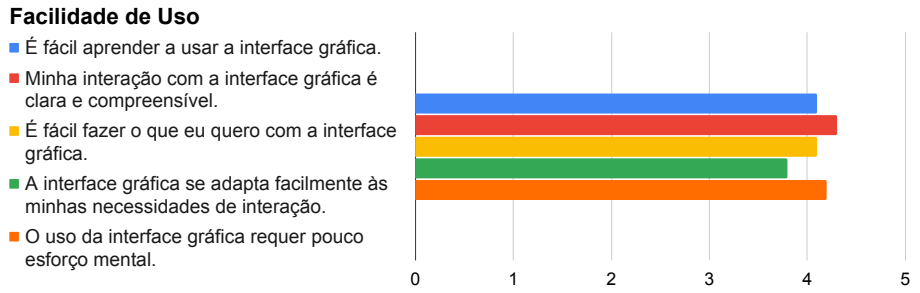
Livre e Esclarecido, seguido por uma lista de instruções detalhadas. Alunos da área de computação da Universidade Federal do Pampa foram convidados a participar, com a participação sendo totalmente voluntária e condicionada à aceitação do termo de consentimento.

3.2. Resultados

Foram obtidas 10 respostas. Os resultados sobre Utilidade Percebida e Facilidade de Uso estão apresentados nas Figuras 3(a) e 3(b), respectivamente. Embora a amostra seja limitada, algumas considerações podem ser feitas. De modo geral, as respostas foram positivas tanto em relação à utilidade quanto à facilidade de uso da interface. Foi observado que a interface gráfica facilita o processo de geração de dados, tornando-o mais intuitivo. Além disso, a maioria dos participantes relatou uma economia de tempo ao utilizar a interface, o que reforça seu valor prático e aplicabilidade no contexto acadêmico.



(a) Utilidade.



(b) Facilidade de Uso.

Figure 3. Resultados da avaliação TAM.

Por outro lado, a interface foi considerada fácil de usar, embora haja espaço para melhorias, como é comum em versões iniciais de ferramentas. A identificação dessas melhorias permite o aprimoramento contínuo da usabilidade e da funcionalidade, atendendo a uma base mais ampla de usuários.

Esperamos que esses resultados, refletindo tanto utilidade quanto facilidade de uso, incentivem maior engajamento dos usuários, especialmente os alunos do Malware DataLab. Além disso, acreditamos que nossa abordagem possa servir de

inspiração para o desenvolvimento de soluções semelhantes em outros ambientes de ensino e aprendizagem voltados para tecnologias emergentes e complexas.

4. Considerações Finais

Neste trabalho, apresentamos uma GUI para o AutoDroid, chamada GUI4MalSynGen, cujo objetivo é tornar a MalSynGen, em sua versão *stand-alone*, mais acessível para os chamados "Hackers do Bem" que se interessam pelo uso de IA generativa no combate a malwares Android. A GUI4MalSynGen simplifica o monitoramento do progresso e desempenho do treinamento das Redes Neurais Artificiais (RNAs), facilitando a interação e minimizando a possibilidade de erros. Com isso, espera-se que o AutoDroid se torne uma solução mais prática para a geração de dados artificiais e o combate a malwares, mesmo para usuários com pouca ou nenhuma experiência com serviços de virtualização e treinamento de redes neurais. Acreditamos que a GUI4MalSynGen terá um impacto positivo no uso de ferramentas de geração de dados, tornando-as mais acessíveis e simplificando sua utilização, com o intuito de alcançar um público ainda maior.

A GUI4MalSynGen será disponibilizada publicamente, integrada ao AutoDroid, permitindo o uso da MalSynGen de forma *stand-alone*. Embora os resultados da avaliação com usuários devam ser interpretados com cautela, devido a certas limitações, o estudo experimental revelou várias oportunidades de melhoria que estão sendo incorporadas no desenvolvimento da GUI para o serviço que será disponibilizado na nuvem como parte do projeto Malware DataLab. Além disso, esperamos que as lições aprendidas no desenvolvimento da GUI4MalSynGen sirvam de inspiração para a criação de novas interfaces voltadas a tecnologias emergentes e complexas, como a IA generativa.

Como trabalhos futuros, planejamos implementar as melhorias identificadas na avaliação TAM, incluindo a tradução do site para o inglês e a ampliação dos parâmetros personalizáveis. Adicionalmente, buscamos agregar novos recursos à versão da solução que será disponibilizada na nuvem. A partir dessa versão online, pretendemos expandir a aplicação do roteiro e do questionário de avaliação para grupos de usuários mais amplos e representativos, incluindo pessoas de diferentes níveis de formação, como o nível técnico, e de diversas regiões do país, com o intuito de aprimorar a análise da interface gráfica e a experiência do usuário. Além disso, é fundamental ampliar a bibliografia sobre o tema, a fim de proporcionar uma base sólida que sustente futuras melhorias e estudos relacionados ao desenvolvimento da solução.

Agradecimentos. A pesquisa contou com apoio parcial da RNP (Programa Hackers do Bem - GT Malware DataLab), da CAPES (Código de Financiamento 001) e da FAPERGS, por meio dos editais 02/2022 (processo 22/2551-0000841-0), 08/2023 e 09/2023 e do termo de outorga 24/2551-0001368-7.

References

Casola, K., Paim, K., Mansilha, R., and Kreutz, D. (2023). DroidAugmentor: uma ferramenta de treinamento e avaliação de cGANs para geração de dados sintéticos.

Davis, F. D., Bagozzi, R. P., and Warshaw, P. R. (1989). User acceptance of computer technology: A comparison of two theoretical models. *Management Science*, 35(8):982–1003.

Fleck, L., Tavares, M. H. F., Eyng, E., Helmann, A. C., and Andrade, M. A. d. M. (2016). Redes neurais artificiais: Princípios básicos. *Revista Eletrônica Científica Inovação e Tecnologia*, 1(13):47–57.

Hu, W. and Tan, Y. (2017). Generating Adversarial Malware Examples for Black-Box Attacks Based on GAN.

Laviola, L., Paim, K., Kreutz, D., and Mansilha, R. (2023). AutoDroid: disponibilizando a ferramenta DroidAugmentor como serviço. In *Anais da XX Escola Regional de Redes de Computadores*, pages 145–150, Porto Alegre, RS, Brasil. SBC.

Likert, R. (1932). A technique for the measurement of attitudes. *Archives of Psychology*, 22(140):1–55.

Miranda, T. C., Gimenez, P.-F., Lalande, J.-F., Tong, V. V. T., and Wilke, P. (2022). Debiasing Android Malware Datasets: How Can I Trust Your Results If Your Dataset Is Biased? *IEEE Transactions on Information Forensics and Security*, 17:2182–2197.

Nogueira, A., Paim, K., Bragança, H., Mansilha, R., and Kreutz, D. (2024a). Geração de dados sintéticos tabulares para detecção de malware android: um estudo de caso. In *Anais do XXIV Simpósio Brasileiro de Segurança da Informação e de Sistemas Computacionais*, pages 808–814, Porto Alegre, RS, Brasil. SBC.

Nogueira, A., Paim, K., Bragança, H., Mansilha, R., and Kreutz, D. (2024b). Malsyngen: redes neurais artificiais na geração de dados tabulares sintéticos para detecção de malware. In *Anais Estendidos do XXIV Simpósio Brasileiro de Segurança da Informação e de Sistemas Computacionais*, pages 129–136, Porto Alegre, RS, Brasil. SBC.