Duta.com : Jurnal Ilmiah Teknologi Informasi dan Komunikasi E-ISSN : 2962 – 3340
Vol 18 No 2 Tahun 2025
https://doi.org/10.47701/dutacom.v18i2.5128

DETEKSI MALWARE ADVERSARIAL PADA JARINGAN IoT: TINJAUAN SISTEMATIS MODEL AI DAN STRATEGI SERANGAN

ADVERSARIAL MALWARE DETECTION IN IoT NETWORKS: A SYSTEMATIC REVIEW OF AI MODELS AND ATTACK STRATEGIES

[1]Andi Novianto, [2]Fatchul Arifin, [3]Herman Dwi Surjono
[1,2,3]Universitas Negeri Yogyakarta
[1*]andinovianto.2023@student.uny.ac.id [2]fatchul@uny.ac.id, [3] hermansurjono@uny.ac.id

Received:	Revised:	Accepted:	Published:
15 July 2025	22 July 2025	23 July 2024	23 August 2025

ABSTRAK

Berkembangnya teknik serangan adversarial malware yang dapat mengelabui sistem AI berbasis DL dan ML telah menarik perhatian para peneliti untuk melakukan pemodelan pengujian serangan terhadap target sistem deteksi malware. Sering kali file malware dianggap sebagai file benign akibat kesalahan deteksi akibat manipulasi data yang dilakukan oleh malware untuk melindungi dirinya. Studi ini menggunakan metodologi tinjauan sistematis terhadap 34 artikel penelitian yang telah difilter berdasarkan aspek inclusion yang secara khusus membahas bagaimana serangan adversarial malware pada jaringan IoT itu dapat dideteksi oleh sistem AI. Tujuan SLR ini adalah menentukan kecenderungan penggunaan jenis AI dalam membangun sistem deteksi malware, memetakan penggunaan algoritma untuk setiap AI, model serangan adversarial malware hingga teknik pengujian yang relevan terhadap metode serangan adversarial tersebut. Hasil kajian ini memperlihatkan bahwa, metode DL dengan algoritma CNN lebih sering dipergunakan untuk membangun sistem deteksi malware secara efektif dibandingkan ML yang dirasakan tidak mampu mengenali jenis varian baru malware. Sedangkan pemodelan serangan cenderung menggunakan metode White Box Based Attacks yang didukung teknik pengujian berbasis Hybrid pada DL.

Kata Kunci : Adversarial Malware Attacks, Malware Detection System, White Box Based Attacks, Black Box Based Attacks, IoT Network Attacks

ABSTRACT

The development of adversarial malware attack techniques that can trick DL and ML-based AI systems has attracted the attention of researchers to conduct attack modeling tests against malware detection system targets. Often, malware files are considered benign files due to detection errors caused by data manipulation carried out by malware to protect itself. This study uses a systematic review methodology of 34 research articles that have been filtered based on the inclusion aspect, which discusses explicitly how adversarial malware attacks on IoT networks can be detected by AI systems. This SLR aims to determine trends in the use of AI types in building malware detection systems and map the use of algorithms for each AI and adversarial malware attack model to relevant testing techniques for these adversarial attack methods. The results of this study show that the DL method with the CNN algorithm is more often used to build effective malware detection systems than ML, which cannot recognize new types of malware variants. Meanwhile, attack modeling uses the White Box Attacks method, supported by Hybrid-based testing techniques on DL.

Keywords*: Adversarial Malware Attacks, Malware Attacks Detection System, White Box Based Attacks, Black Box Based Attacks, IoT Network Attacks*

PENDAHULUAN

Sebagai salah satu ancaman serius bagi keamanan informasi, malicious code atau malware dibuat dengan tujuan untuk menginfeksi komputer, merusaknya dan mencuri data penting(R. Ali et al., 2022), kini telah dilengkapi kemampuan canggih yang mampu mengelabui program pendeteksi malware seperti antivirus dan program sejenisnya meski telah dilengkapi fitur artificial intelegence seperti Deep Learning dan Machine Learning (H. Li et al., 2023). Kejahatan terbesar yang diakibatkan oleh malware masih didominasi Stealer/Spyware/Keylogger yang bertujuan untuk pembobolan dan pencurian data (Sophos, 2024) mengakibatkan kerugian besar dalam bidang finansial, kesehatan dan beberapa perangkat IoT. Jenis malware sendiri dapat dibedakan menjadi beberapa macam, antara lain spyware, worms, backdoor, adware, virus, keyloggers, trojan horse, rootkits, crytojacking, ransomware, botnet, fileless malware dan obfuscated malware (Maniriho et al., 2024) (Yan et al., 2023). Perkembangan jenis baru varian malware seperti trojan RAT, dapat dilihat dampaknya dengan semakin meningkatnya trafik anomali pada jaringan yang mengakibatkan menurunnya perfoma jaringan internet, rentannya keamanan siber (BSSN, 2023) hingga munculnya permasalahan baru pada jaringan Internet-of-Things (IoT) (Sánchez Sánchez et al., 2024).

Banyak penelitian yang mengusulkan metode pendeteksi malware, seperti Light Gradient Boosting (X. , L. X. , W. F. , L. W. , L. A. Li, 2021), analisis fitur bytes terhadap malware berbasis ML(Priyadarshan, 2021) hingga metode berbasis LSTM untuk mengklasifikasi jenis malware menggunakan sistem calling (Or-Meir, 2021). Metode Machine Learning berbasis algoritma Random Forest menunjukkan persentase penggunaan terbanyak dalam sistem deteksi malware pada mesin berbasis Windows(Maniriho et al., 2024; Pascal Maniriho et al., 2023). Kelemahan metode ini adalah ketidak mampuannya dalam mendeteksi serta mengklasifikasikan jenis malware baru, sehingga membutuhkan dataset besar sebagai data training(JARETH A.V., 2023). Sedangkan metode Deep Learning (DL) cenderung bekerja efektif dalam mengenali varian baru bahkan mengklasifikasikannya (John S.A., 2022), sehingga banyak studi yang bergeser menggunakan DL sebagai basis metode pendeteksi khususnya menggunakan algoritma CNN(Maniriho et al., 2024).

Dengan memanfaatkan AI seperti ML dan DL, sistem pendeteksi diharapkan mampu menangani masalah serangan siber dan melakukan investigasi pada keamanan aplikasi seperti mendeteksi malware, mendeteksi spam, mendeteksi network anomaly (Alatwi & Morisset, 2021) dan mencegah serangan pada jaringan IoT (Sánchez Sánchez et al., 2024). Namun, dengan berkembangnya teknik serangan adversarial, malware lebih sulit terdeteksi karena memanipulasi nilai input terhadap model guna mengelabui sistem pendeteksi (Maniriho et al., 2024). Setelah berhasil menginfeksi dan mencuri data, malware akan mentransmisikannya ke server peretas atau mengirimkan data berbahaya ke komputer korban lainnya (DdoS) sehingga menurunkan perfoma kerja jaringan, sebagai contoh menggunakan metode the threat of AEs (Alatwi & Morisset, 2021). Minimnya penelitian yang mengulas bagaimana melakukan identifikasi dan analisis serangan adversial malware pada jaringan IoT, mendorong kami untuk melakukan pendekatan Studi Literatur Review dengan mengumpulkan dan mengintegrasikan beberapa kajian penelitian yang ada terkait masalah tersebut. Tujuan utama dalam kajian ini adalah mengumpulkan informasi tentang jenis metode AI apa saja yang digunakan, mengidentifikasi jenis algoritma yang diterapkan dalam sistem pendeteksi, menentukan dan menganalisis model serangan adversarial maupun penggunaan teknik analisis malware yang sering digunakan dalam kajian penelitian. Kontribusi utama SLR ini adalah memetakan penerapan AI dan metode deteksi serangan adversarial yang dilakukan malware sehingga mampu memberikan pemahaman tentang efektifitas sistem deteksi yang digunakan.

METODE

1. Pertanyaan Dalam Penelitian

Daftar pertanyaan penelitian atau RQs telah disusun untuk menentukan arah dan tujuan utama dalam kajian literatur review ini. Rancangan entitas dalam kajian didesain menggunakan sistem *Population*, *Intervention*, *Comparation*, *Outcomes and Context* (PICOC) seperti yang dijelaskan oleh Kitchenham and Charters (Kitchenham & Charters, 2007). Susunan pertanyaan dalam penelitian berbasis PICOC dapat dilihat dalam tabel 1.

Tabel 1. PICOC

PICOC	Deskripsi
Populasi (*Population*)	Kumpulan data malware, Jaringan IoT, mesin berbasis windows, sistem android dan sistem komputer
Intervensi (*Intervention*)	Deteksi serangan malware adversarial menggunakan kecerdasan buatan, deteksi malware, deteksi serangan pada jaringan IoT, deteksi serangan botnet.
Perbandingan (*Comparison*)	Perbandingan sistem deteksi serangan malware tanpa metode serangan adversarial

Keluaran (*Outcomes*)	Efektivitas dalam mendeteksi serangan adversarial pada malware
Konteks (*Context*)	Implementasi teknik deteksi malware di lingkungan perusahaan skala besar

Kajian dalam SLR ini dibuat berdasarkan empat pertanyaan penelitian (RQs) yang berkaitan dengan Adversarial Malware attack in IoT Network. Pertanyaan-pertanyaan tersebut digunakan untuk mengidentifikasi beberapa literatur yang berhubungan dengan masalah tersebut, yang mencakup beberapa hal sebagai berikut :
RQ1 : Jenis metode artificial intelengence apakah yang paling sering digunakan untuk mendeteksi serangan adverserial malware dalam jaringan IoT?
RQ2 : Jenis algoritma apakah yang paling banyak digunakan dalam sistem deteksi tersebut?
RQ3 : Jenis model serangan adversarial apakah yang sering dijadikan eksperimen?
RQ4 : Jenis metode pengujian apakah yang sering digunakan peneliti untuk mengidentifikasi dan mengklasifikasi jenis malware?

2. Kriteria Inklusi dan Ekslusi

Tabel 2 berikut ini menjelaskan kriteria inklusi dan eksklusi dalam pembuatan SLR tentang Adversarial Malware Attack Detection System in IoT Networks.

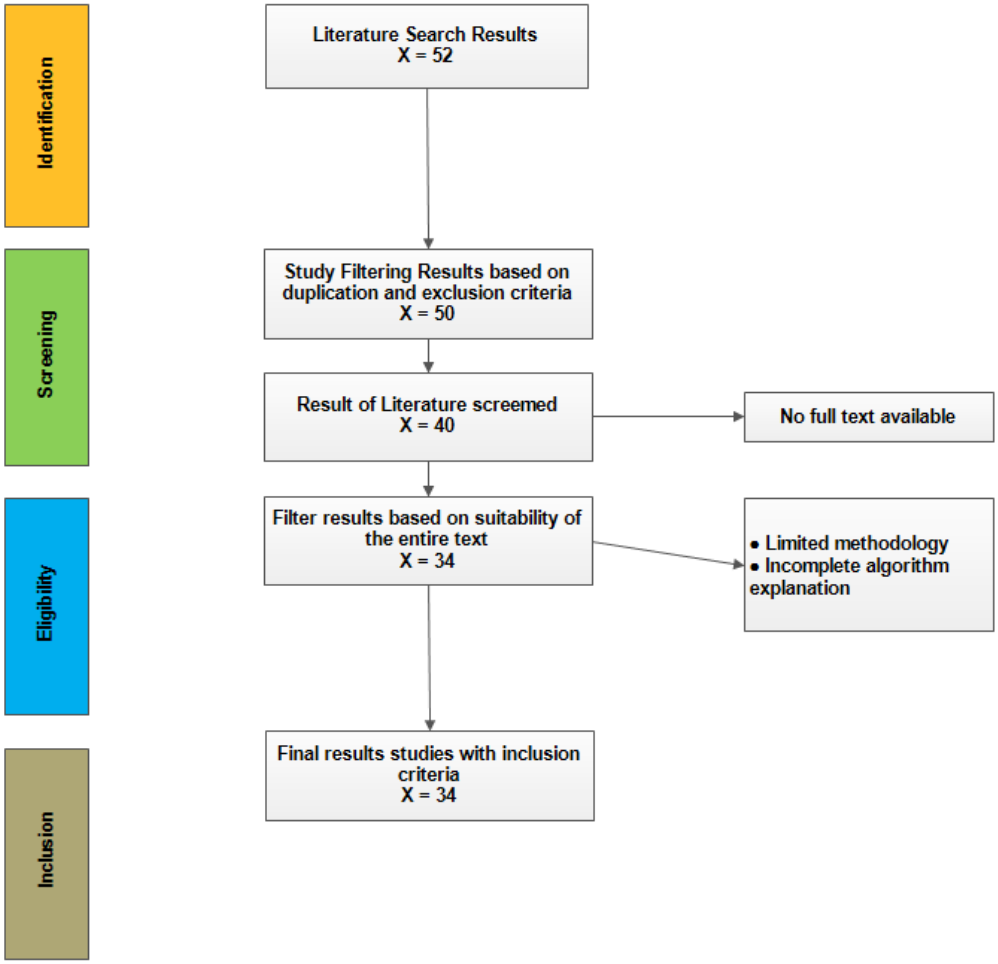
Tabel 2. Jenis Kriteria Inklusi dan Eksklusi

Kriteria	Inklusi	Eksklusi
Jenis Kasus	Deteksi Serangan Adversarial Malware dalam Jaringan IoT	Tidak Termasuk Kategori Deteksi Serangan Adversarial Malware dalam Jaringan IoT
Bahasa	Inggris	Selain Bahasa Inggris
Tanggal Publikasi	2017 - 2024	Sebelum 2017
Jenis Publikasi	Terakreditasi Scopus atau Terindek Scopus	Referensi abu-abu

3. Sumber Data dan Literasi Riset

Dengan menggunakan keyword Adversarial Malware Attack Detection System in IoT Network pada scopus menghasilkan 52 kajian yang layak untuk dieksplor. Investigasi dan analisis berbagai literatur tersebut difokuskan pada kajian tentang bagaimana menguji adversarial malware yang berpotensi melakukan serangan pada jaringan IoT.

https://doi.org/10.47701/dutacom.v18i2.5128



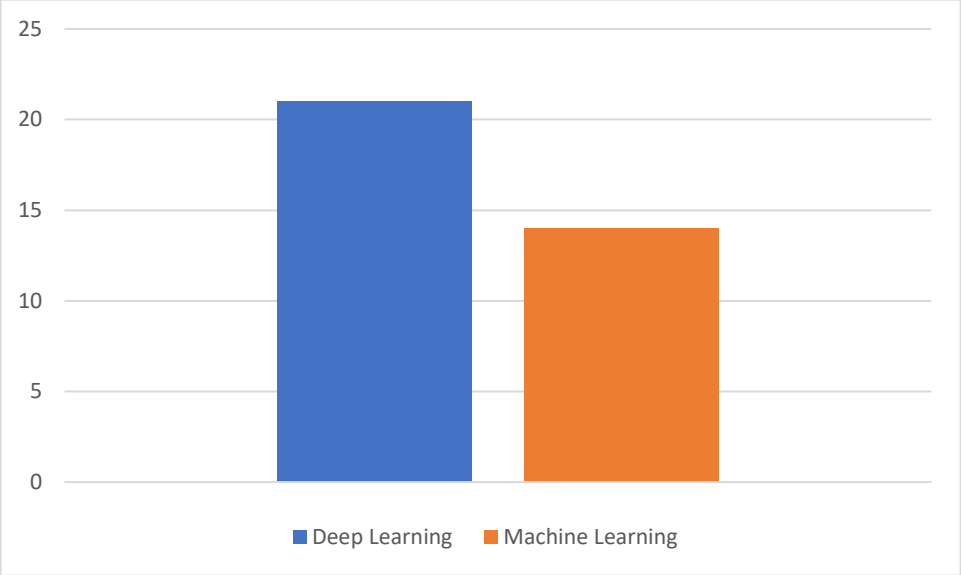
Gambar 1. Metodologi PRISMA

Hasil Penelitian

1. Metode AI Dalam Mendeteksi Malware (RQ1)

Teknologi ML dan DL sering digunakan para attacker untuk memperbarui dan meningkatkan kemampuan malware dalam melakukan serangan (Maniriho et al., 2024). Attacker akan memanfaatkan inputan palsu (adversarial sample) untuk mengelabui machine learning model sehingga menghasilkan keputusan yang salah terhadap serangan tersebut (T. Ali et al., 2023). Biggio et al. (Biggio et al., 2013) merupakan peneliti pertama yang mendesain metode gradient-based untuk menghasilkan adversarial samples yang dapat dianalisis menggunakan metode linear classifiers, SVM (Support Vector Machines) dan neural networks (Yang & Yin, 2023). Machine Learning memiliki perfoma baik dalam mendeteksi malware selama memiliki data training yang lengkap, namun kurang efektif ketika mengenali varian baru malware dan rentan terhadap serangan adversarial (Yang & Yin, 2023). Sedangkan DL lebih responsif dalam mengenali jenis malware baru meski belum memiliki dataset yang menjadi pola training (Maniriho et al., 2024), dengan cara mengekstrak pola yang sesuai dengan data berdasarkan arsitektur jaringan syaraf yang lebih canggih(Aslan O & Yilmaz A.A., 2021). Dalam hal mendeteksi dan mengklasifikasi jenis malware secara tradisional, teknik DL cenderung lebih sering dipergunakan dalam berbagai publikasi ilmiah dibandingkan model ML (Maniriho et al., 2024). Meski demikian, dalam percobaan generate adversarial samples yang diuji menggunakan anti virus berbasis AI seperti Blackberry Cyclance menyatakan malware tersebut sebagai benign file(Skylight cyber, 2019) (Maniriho et al., 2024). Beberapa studi mengusulkan framework baru untuk mendeteksi serangan adversarial malware dalam jaringan IoT berbasis DL, seperti metode MalConv dan GAN (Sabuhi et al., 2021; Yang & Yin, 2023), strategi GMA (Grey Box Malware Agent) (Rathore et al., 2022), metode robustnes pada perangkat android (Rathore, Sahay, et al., 2021), sistem DGCNN dan Reinforcement Learning (Chen et al., 2023), metode manipulasi biner (F. Wang et al., 2021), strategi Adversarial attack detection framework based on optimized weighted conditional stepwise adversarial network (Barik et al., 2024), penerapan algoritma CNN (Zhang et al., 2024)

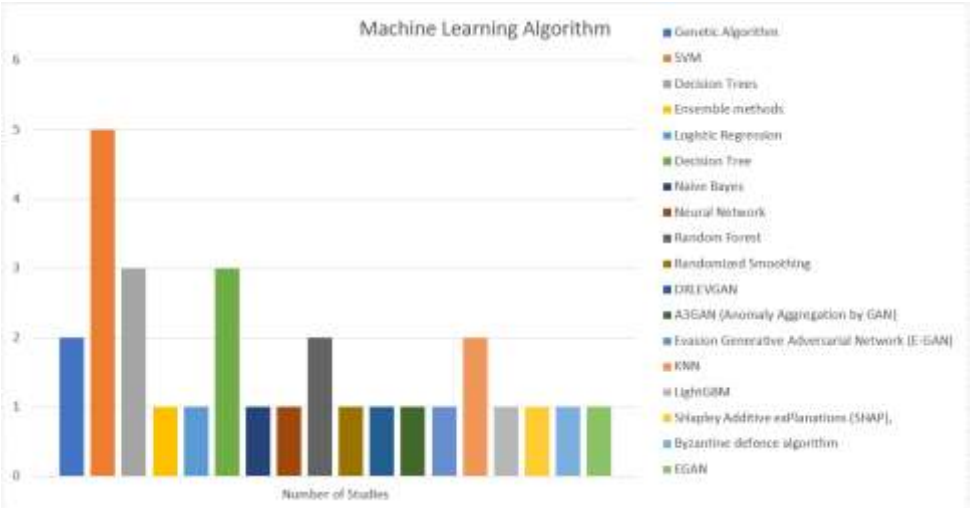
(Rathore, Bandwala, et al., 2021)(Rust-Nguyen et al., 2023) (Moti et al., 2021) (Ijas et al., 2021) dan DNN (Barik et al., 2024), implementasi teknik AMGmal (Zhan et al., 2023), dynamic analysis berbasis neural network (Stokes et al., 2017), framework GAPGAN (J. Yuan et al., 2020), kombinasi metode FCG, GCN (Lu et al., 2024) dan FCG (H. Li et al., 2023), OWCSAN (Barik et al., 2024), EGAN (Saravanan et al., 2023). Sedangkan perbaikan metode Machine Learning yang telah diadaptasi agar dapat mengenali jenis serangan adversarial telah dibahas dalam beberapa studi seperti penerapan algoritma genetic (P. Yuan et al., 2023) (Liu et al., 2019), penggunaan teknik obfuscation code (J. Wang et al., 2021), ensemble methods (Rathore, Sahay, et al., 2021), DRLEVGAN Training Algorithm (Anand et al., 2023), penerapan teknik SVM (Taheri et al., 2020) (Taheri et al., 2020) (Woessner, 2020), metode logistic regression (Dinakarrao et al., 2019), randomized smoothing (Gibert et al., 2024), backdoor classification (Reddy & Lakshmi, 2021), teknik A3GAN (Taheri et al., 2021) dan teknik E-GAN (Debicha et al., 2023). Dalam periode penelitian mulai tahun 2017 hingga Mei 2024 menunjukkan kecenderungan penggunaan metode Deep Learning sebagai metode pendeteksi serangan adversarial malware.



Gambar 2. Banyaknya Riset Berbasis Kecerdasan Buatan (AI)

2. Penggunaan Algoritma Dalam Sistem Deteksi (RQ2)

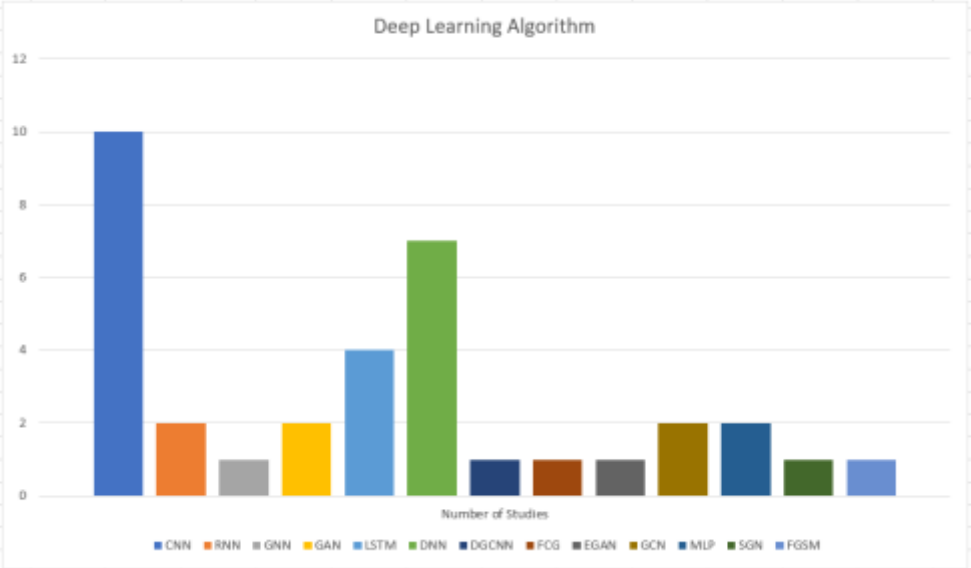
Hasil kajian sebelumnya menunjukkan bahwa 60% studi cenderung menggunakan DL sebagai basis metode pendeteksi serangan adversarial dibandingkan ML. oleh karena itu perlu diidentifikasi dan dianalisis seberapa banyaknya jenis algoritma yang paling sering dipakai dalam riset baik menggunakan DL atau ML untuk memetakan tren penggunaan algoritma yang paling umum dalam membangun sistem pendeteksi serangan adversarial malware khususnya dalam jaringan IoT. Berikut ini adalah hasil pengelompokkan banyaknya algoritma yang digunakan dalam studi penelitian berdasarkan metode DL dan ML.



Gambar 3. Jumlah Indikasi Penggunaan Algoritma Berbasis ML

https://doi.org/10.47701/dutacom.v18i2.5128

Dari gambar 3 tersebut memperlihatkan bahwa algoritma SVM paling sering digunakan dalam penelitian untuk membangun sistem deteksi serangan adversarial malware berbasis ML.



Gambar 4. Jumlah Indikasi Penggunaan Algoritma Berbasis DL

Sedangkan gambar 4, memperlihatkan penggunaan algoritma CNN menduduki peringkat terbanyak dalam studi berbasis DL untuk membangun sistem deteksi serangan adversarial malware.

3. Model Serangan Adversarial Malware (RQ3)

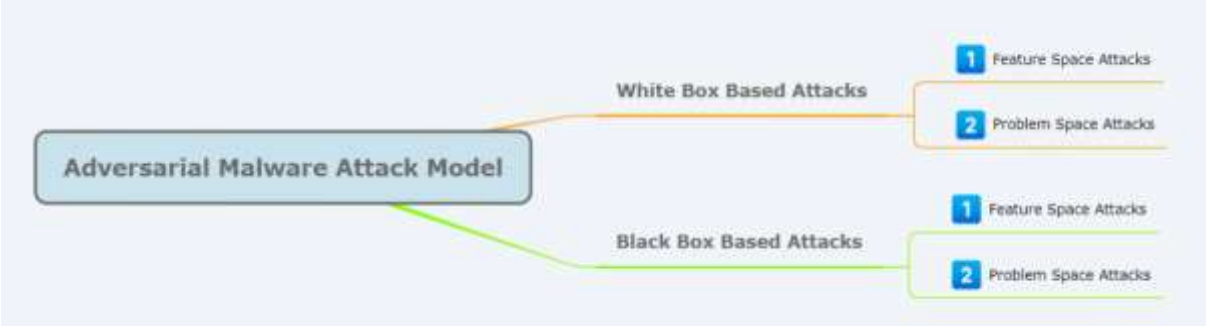
Tujuan utama serangan adversarial adalah menggenerate sampel serangan yang dilakukan malware agar dapat mengelabui model deteksi malware sehingga malware berbahaya dianggap sebagai benign file (Debicha et al., 2023). Teknik serangan adversarial dapat diklasifikasikan menjadi dua jenis, yaitu white box-based attacks dan black box-based attacks (Martins et al., 2020).

a. White box-based attacks
   Model serangan dilakukan terhadap model deteksi berdasarkan informasi tentang struktur dan parameter model (Alatwi & Morisset, 2021; Maniriho et al., 2024).
b. Black box-based attacks
   Serangan dilakukan tanpa bekal informasi tentang model deteksi, sehingga pengubahan nilai input pada model disesuaikan output yang ditampilkan guna mengelabui model deteksi(Alatwi & Morisset, 2021; F. Wang et al., 2021; J. Yuan et al., 2020).

Kedua model serangan tersebut dalam pekerjaannya dapat dikelompokkan dalam dua tipe berdasarkan basis pengetahuan serangan terhadap target model deteksi (sesuai gambar 5), yaitu:

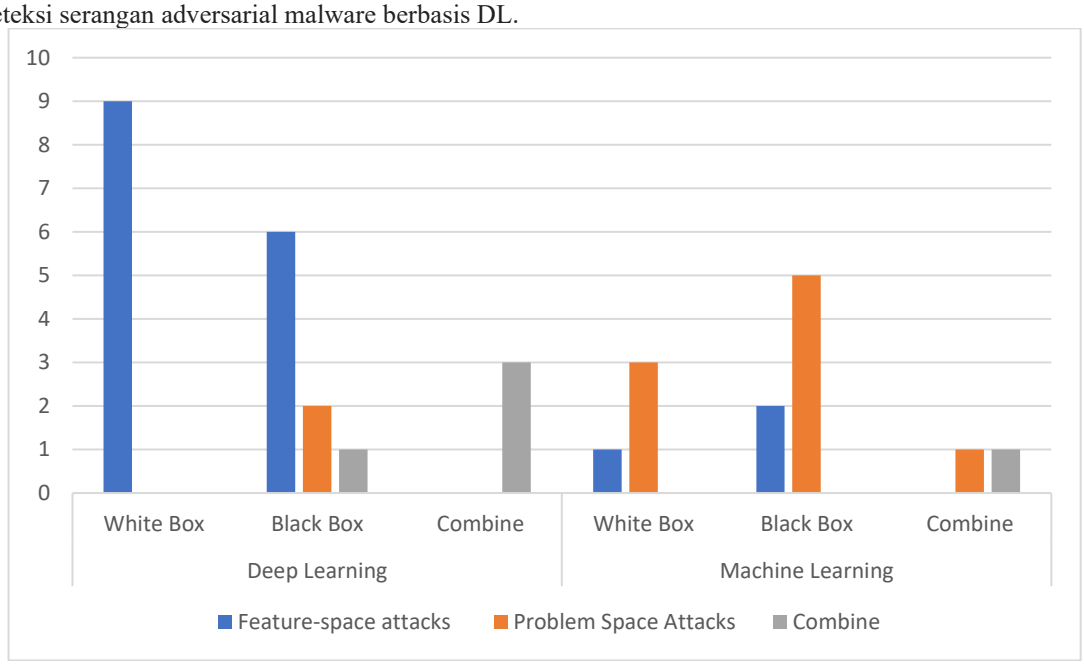
a. Feature-space attacks merupakan teknik memanipulasi dataset training yang dapat dilakukan dengan menyisipkan data berbahaya untuk menurunkan kinerja model (data poisoning attacks) atau memanipulasi label dalam dataset training sehingga mengaburkan analisis malware (Alatwi & Morisset, 2021).
b. Problem-space attacks merupakan teknik manipulasi karakteristik asli data tanpa mengubah label, yang dapat digenerate menggunakan metode Generative Adversarial (GAN) (Yang & Yin, 2023)(Anand et al., 2023), A3GAN (Taheri et al., 2021), Reinforcement Learning (Chen et al., 2023) dan metode Gradien Based Attacks(Ijas et al., 2021).

https://doi.org/10.47701/dutacom.v18i2.5128



Gambar 5. Jenis Model Serangan Adversarial Malware

Pada gambar 6, menunjukkan bahwa model serangan White box based sering dipergunakan pada teknik deteksi serangan adversarial malware berbasis DL.



Gambar 6. Perbandingan Penggunaan Metode Analisis Serangan Adversarial Malware Berbasis Dl dan ML

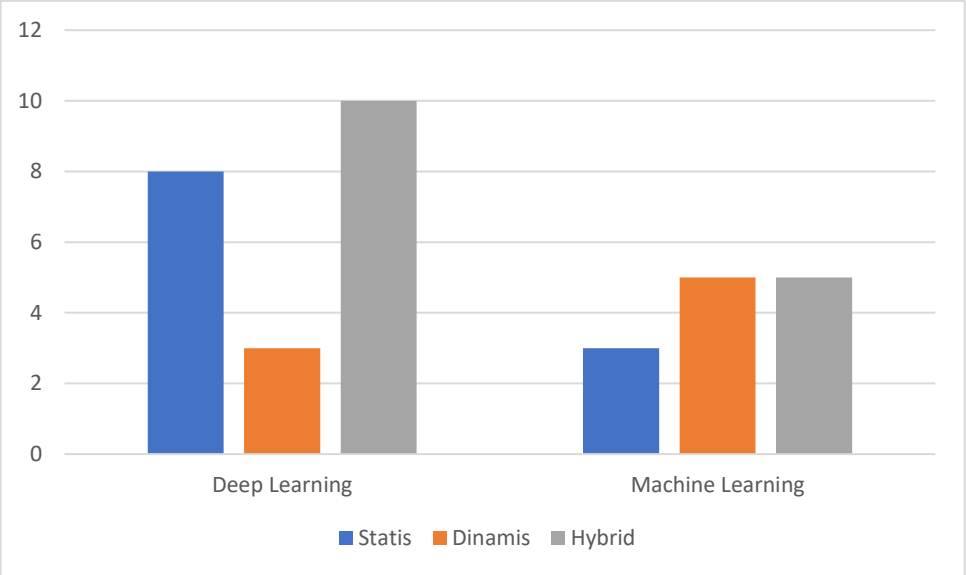
4. Teknik Pengujian Serangan Adversarial Malware (RQ4)

Teknik deteksi malware yang dilakukan secara langsung pada file malware yang belum dilengkapi kemampuan adversarial, dikelompokkan dalam empat jenis yaitu statis, dinamik, memory dan hybrid (Maniriho et al., 2024).

a. Static analysis dilakukan dengan memanfaatkan basis data signature yang bersifat tetap (Carlin D et al., 2019) kemudian dicocokkan terhadap kode berbahaya dalam malware. Metode ini sering digunakan program anti virus yang telah dilengkapi dengan database signature.
b. Dynamic analysis merupakan teknik deteksi berbasis behaviour yang dilakukan dengan menganalisis pola perilaku malware yang terindikasi melakukan perubahan data pada file sistem atau bahkan memodifikasi alamat jaringan (Li C et al., 2022). Dengan menandai perilaku aneh yang berpotensi membahayakan sistem, akan mempercepat proses deteksi meski membutuhkan resource besar dan menghasilkan false positive tinggi (Maniriho et al., 2024) (J. Wang et al., 2021).
c. Memory Analysis, dilakukan dengan mengcapture semua aktifitas malware yang terekam dalam memory, mulai dari proses, kondisi jaringan hingga proses injeksi sistem (Maniriho et al., 2024).
d. Teknik deteksi hybrid merupakan kombinasi dari metode sebelumnya yaitu statis dan dinamis (Rathore et al., 2022).

Dalam kajian literatur ini, telah dilakukan identifikasi dalam tiap studi apakah dalam proses melakukan analisis malware menggunakan teknik statik, dinamik atau menggunakan teknik hybrid.

https://doi.org/10.47701/dutacom.v18i2.5128



Gambar 7. Indikasi Banyaknya Penggunaan Metode Analisis Malware

Dalam gambar 7, menunjukkan bahwa metode hybrid lebih banyak digunakan dalam teknik analisis berbasis Deep Learning karena keunggulan dalam efektifitas mendeteksi varian baru malware khususnya serangan adversarial. Pada perangkat IoT seperti Android, teknik hybrid relatif efektif jika dikombinasikan dengan sistem FCG meski membutuhkan resource besar. Dari 34 kajian studi menunjukkan, hanya satu riset yang mensimulasikan serangan adversarial pada network yang ditujukan pada sistem NIDS menggunakan metode ML.

ISU DAN TANTANGAN

Meskipun data SLR memperlihatkan keseimbangan penggunaan metode DL dan ML dalam mendeteksi serangan adversarial malware, namun ada beberapa isu dan tantangan yang dapat dijadikan pertimbangan dalam penelitian berikutnya, antara lain:

a. Metode Machine Learning masih digunakan untuk mendeteksi malware pada jaringan IoT, namun harus selalu diupdate dataset dan informasi properti varian malware terbaru agar lebih optimal dalam mengidentifikasi dan menganalisis tipe malware baru (Moti et al., 2021).
b. Teknik hybrid yang diterapkan pada android, akan lebih efisien jika menggunakan FCG (Function Call Graph) meski menguras sumber daya komputer karena memuat banyak node API (Lu et al., 2024) sehingga sistem ini perlu dikembangkan lebih lanjut.
c. Perlunya pemodelan dan generate malware yang mampu melakukan uji serangan adversarial, khususnya evasion attacks yang disimulasikan pada NIDS (Network Intrusion Detection System) menggunakan botnet (Debicha et al., 2023).

KESIMPULAN

Dari hasil analisis SLR yang telah dilakukan, memperlihatkan bahwa rata-rata studi tentang pengujian serangan adversarial malware cenderung menggunakan sistem deteksi berbasis DL dengan alasan kemampuan identifikasi dan analisis lebih baik ketika menemukan varian baru malware dibandingkan ML. Dalam literasi juga menunjukkan bahwa algoritma CNN (*Convolutional Neural Network*) lebih sering dipergunakan dalam metode deteksi DL dengan menerapkan model serangan adversarial berbasis White Box. Sedangkan algoritma SVM (*Support Vector Machine*) mendominasi dalam metode deteksi berbasis ML dengan pengujian serangan adversarial berbasis Black Box. Efektifitas deteksi malware menggunakan hybrid dapat dikombinasikan dengan sistem FCG, namun harus memperhatikan kekurangan dalam hal penggunaan resource hardware yang terlalu besar. Di samping itu, minimnya informasi tentang konsep White Box based attacks dan Black Box Based Attacks yang disimulasikan dalam serangan jaringan IoT hasil dari generate malware berkemampuan adversarial attacks menjadi tantangan baru pada penelitian berikutnya.

DAFTAR PUSTAKA

Alatwi, H. A., & Morisset, C. (2021). *Adversarial Machine Learning In Network Intrusion Detection Domain: A Systematic Review*. http://arxiv.org/abs/2112.03315

Ali, R., Ali, A., Iqbal, F., Hussain, M., & Ullah, F. (2022). Deep Learning Methods for Malware and Intrusion Detection: A Systematic Literature Review. In *Security and Communication Networks* (Vol. 2022). Hindawi Limited. https://doi.org/10.1155/2022/2959222

Ali, T., Eleyan, A., & Bejaoui, T. (2023). Detecting Conventional and Adversarial Attacks Using Deep Learning Techniques: A Systematic Review. *2023 International Symposium on Networks, Computers and Communications, ISNCC 2023*. https://doi.org/10.1109/ISNCC58260.2023.10323872

Anand, K., Rao Budaraju, R., Kumar, S., Rao, B. M., & Sah, B. (2023). Evasion-Aware Botnet Attack Detection using Deep Reinforcement Adversarial Learning. In *Original Research Paper International Journal of Intelligent Systems and Applications in Engineering IJISAE* (Vol. 2024, Issue 5s). www.ijisae.org

Aslan O, & Yilmaz A.A. (2021). A new malware classification framework based on deep learning algorithms. *IEEE Access*. https://ieeexplore.ieee.org/document/9455368

Barik, K., Misra, S., & Fernandez-Sanz, L. (2024). Adversarial attack detection framework based on optimized weighted conditional stepwise adversarial network. *International Journal of Information Security*. https://doi.org/10.1007/s10207-024-00844-w

Biggio, B., Corona, I., Maiorca, D., Nelson, B., Nedimˇ, N., Nedimšrndi´c, N., Laskov, P., Giacinto, G., & Roli, F. (2013). *LNAI 8190 - Evasion Attacks against Machine Learning at Test Time*. https://link.springer.com/chapter/10.1007/978-3-642-40994-3_25

BSSN. (2023). *Lanskap Keamanan Siber Indonesia*.

Carlin D, O'Kane P, & Sezer S. (2019). A cost analysis of machine learning using dynamic runtime opcodes for malware detection. *Computers & Security*, *85*, 138–155.

Chen, Y., Feng, Y., Wang, Z., Zhao, J., Wang, C., & Liu, Q. (2023). IMaler: An Adversarial Attack Framework to Obfuscate Malware Structure Against DGCNN-Based Classifier via Reinforcement Learning. *IEEE International Conference on Communications*, *2023-May*, 790–796. https://doi.org/10.1109/ICC45041.2023.10279372

Debicha, I., Cochez, B., Kenaza, T., Debatty, T., Dricot, J. M., & Mees, W. (2023). Adv-Bot: Realistic adversarial botnet attacks against network intrusion detection systems. *Computers and Security*, *129*. https://doi.org/10.1016/j.cose.2023.103176

Dinakarrao, S. M. P., Amberkar, S., Bhat, S., Dhavlle, A., Sayadi, H., Sasan, A., Homayoun, H., & Rafatirad, S. (2019, June 2). Adversarial attack on microarchitectural events based malware detectors. *Proceedings - Design Automation Conference*. https://doi.org/10.1145/3316781.3317762

Gibert, D., Demetrio, L., Zizzo, G., Le, Q., Planes, J., & Biggio, B. (2024). *Certified Adversarial Robustness of Machine Learning-based Malware Detectors via (De)Randomized Smoothing*. http://arxiv.org/abs/2405.00392

Ijas, A. H., Vinod, P., Zemmari, A., Harikrishnan, D., Poulose, G., Jose, D., Mercaldo, F., Martinelli, F., & Santone, A. (2021). Vulnerability evaluation of android malware detectors against adversarial examples. *Procedia Computer Science*, *192*, 3320–3331. https://doi.org/10.1016/j.procs.2021.09.105

JARETH A.V. (2023). *The pros, cons and limitations of AI and machine learning in antivirus software - Emsisoft — security blog*. Https://Www.Emsisoft.Com/En/Blog/35668/the-Pros-Cons-and-Limitations-of-Ai-and-Machine-Learning-in-Antivirus-Software/. https://www.emsisoft.com/en/blog/35668/the-pros-cons-and-limitations-of-ai-and-machine-learning-in-antivirus-software/

John S.A. (2022). *95% of new malware threats target windows OS*. Https://Www.Dailyhostnews.Com/Malware-Threats-Aimed-at-Windows. https://www.sciencedirect.com/science/article/pii/S0164121223003163#b119

Kitchenham, B., & Charters, S. (2007). *Guidelines for performing Systematic Literature Reviews in Software Engineering*.

Li C, Lv Q, Li N, Wang Y, Sun D, & Qiao Y. (2022). A novel deep framework for dynamic malware detection based on API sequence intrinsic features. *Computers & Security*, *116*.

Li, H., Cheng, Z., Wu, B., Yuan, L., Gao, C., Yuan, W., & Luo, X. (2023). *Black-box Adversarial Example Attack towards FCG Based Android Malware Detection under Incomplete Feature Information*.

Li, X. , L. X. , W. F. , L. W. , L. A. (2021). A Malware Detection Method Based on Machine Learning and Ensemble of Regression Trees. In: 2021 2nd. *International Conference on Artificial Intelligence and Information Systems. Pp. 1–6. Google Scholar*, 1–6.

Liu, X., Du, X., Zhang, X., Zhu, Q., Wang, H., & Guizani, M. (2019). Adversarial samples on android malware detection systems for IoT systems. *Sensors (Switzerland)*, *19*(4). https://doi.org/10.3390/s19040974

Lu, X., Zhao, J., Zhu, S., & Lio, P. (2024). SNDGCN: Robust Android malware detection based on subgraph network and denoising GCN network. *Expert Systems with Applications*, *250*. https://doi.org/10.1016/j.eswa.2024.123922

Maniriho, P., Mahmood, A. N., & Chowdhury, M. J. M. (2024). A systematic literature review on Windows malware detection: Techniques, research issues, and future directions. *Journal of Systems and Software*, *209*. https://doi.org/10.1016/j.jss.2023.111921

Martins, N., Cruz, J. M., Cruz, T., & Henriques Abreu, P. (2020). Adversarial Machine Learning Applied to Intrusion and Malware Scenarios: A Systematic Review. In *IEEE Access* (Vol. 8, pp. 35403–35419). Institute of Electrical and Electronics Engineers Inc. https://doi.org/10.1109/ACCESS.2020.2974752

Moti, Z., Hashemi, S., Karimipour, H., Dehghantanha, A., Jahromi, A. N., Abdi, L., & Alavi, F. (2021). Generative adversarial network to detect unseen Internet of Things malware. *Ad Hoc Networks*, *122*. https://doi.org/10.1016/j.adhoc.2021.102591

Or-Meir, O. , C. A. , E. Y. , R. L. , N. N. (2021). Pay Attention: Improving Classification of PE Malware Using Attention Mechanisms Based on System Call Analysis. *International Joint Conference on Neural Networks. IJCNN*, 1–8.

Pascal Maniriho, Abdun Naser Mahmood, & Mohammad Jabed Morshed Chowdhury. (2023). A systematic literature review on windows malware detection: Techniques, research issues, and future directions. *Journal of Systems and Software*.

Priyadarshan, P. , S. P. , R. A. , P. (2021). Machine Learning Based Improved Malware Detection Schemes. In: 2021 11th International Conference on Cloud Computing. *Data Science Engineering (Confluence)*, 925–931.

Rathore, H., Bandwala, T., Sahay, S. K., & Sewak, M. (2021). Poster Abstract: Are CNN based Malware Detection Models Robust?: Developing Superior Models using Adversarial Attack and Defense. *SenSys 2021 - Proceedings of the 2021 19th ACM Conference on Embedded Networked Sensor Systems*, 355–356. https://doi.org/10.1145/3485730.3492867

Rathore, H., Sahay, S. K., Dhillon, J., & Sewak, M. (2021). Designing Adversarial Attack and Defence for Robust Android Malware Detection Models. *Proceedings - 51st Annual IEEE/IFIP International Conference on Dependable Systems and Networks - Supplemental Volume, DSN-S 2021*, 29–32. https://doi.org/10.1109/DSN-S52858.2021.00025

Rathore, H., Samavedhi, A., Sahay, S. K., & Sewak, M. (2022). Are Malware Detection Models Adversarial Robust Against Evasion Attack? *INFOCOM WKSHPS 2022 - IEEE Conference on Computer Communications Workshops*. https://doi.org/10.1109/INFOCOMWKSHPS54753.2022.9798221

Reddy, G. S., & Lakshmi, S. M. (2021). Retraction: Exploring adversarial attacks against malware classifiers in the backdoor poisoning attack (IOP Conf. Ser.: Mater. Sci. Eng. 1022 012037). *IOP Conference Series: Materials Science and Engineering*, *1022*(1), 012125. https://doi.org/10.1088/1757-899x/1022/1/012125

Rust-Nguyen, N., Sharma, S., & Stamp, M. (2023). Darknet traffic classification and adversarial attacks using machine learning. *Computers and Security*, *127*. https://doi.org/10.1016/j.cose.2023.103098

Sabuhi, M., Zhou, M., Bezemer, C. P., & Musilek, P. (2021). Applications of Generative Adversarial Networks in Anomaly Detection: A Systematic Literature Review. In *IEEE Access* (Vol. 9, pp. 161003–161029). Institute of Electrical and Electronics Engineers Inc. https://doi.org/10.1109/ACCESS.2021.3131949

Sánchez Sánchez, P. M., Huertas Celdrán, A., Bovet, G., & Martínez Pérez, G. (2024). Adversarial attacks and defenses on ML- and hardware-based IoT device fingerprinting and identification. *Future Generation Computer Systems*, *152*, 30–42. https://doi.org/10.1016/j.future.2023.10.011

Saravanan, T., Deepa, S., & Sasikumar, P. (2023). Advanced EGAN-IDS Framework for Resilience against Adversarial Attacks using Multi-headed Attention Module. *Procedia Computer Science*, *230*, 203–213. https://doi.org/10.1016/j.procs.2023.12.075

Skylight cyber. (2019). *Skylight cyber* . Https://Skylightcyber.Com/2019/07/18/Cylance-i-Kill-You/.

Sophos. (2024). *Sophos 2024 Threat Report: Cybercrime on Main Street*.

Stokes, J. W., Wang, D., Marinescu, M., Marino, M., & Bussone, B. (2017). *Attack and Defense of Dynamic Analysis-Based, Adversarial Neural Malware Detection Models*.

Taheri, R., Javidan, R., Shojafar, M., Pooranian, Z., Miri, A., & Conti, M. (2020). On defending against label flipping attacks on malware detection systems. *Neural Computing and Applications*, *32*(18), 14781–14800. https://doi.org/10.1007/s00521-020-04831-9

Taheri, R., Shojafar, M., Alazab, M., & Tafazolli, R. (2021). Fed-IIoT: A Robust Federated Malware Detection Architecture in Industrial IoT. *IEEE Transactions on Industrial Informatics*, *17*(12), 8442–8452. https://doi.org/10.1109/TII.2020.3043458

Wang, F., Lu, Y., Wang, C., & Li, Q. (2021). Binary Black-Box Adversarial Attacks with Evolutionary Learning against IoT Malware Detection. *Wireless Communications and Mobile Computing*, *2021*. https://doi.org/10.1155/2021/8736946

Wang, J., Yang, T., Yao, P., Yan, B., Hao, W., & Yang, Q. (2021). Adversarial Malware Examples for Terminal Cyberspace Attack Analysis in Cyber-Physical Power Systems. *Proceedings - 2021 International Conference on Power System Technology: Carbon Neutrality and New Type of Power System, POWERCON 2021*, 1865–1870. https://doi.org/10.1109/POWERCON53785.2021.9697702

Woessner, P. A. (2020). *Adversarial Attack Prevention And Malware Detection System*.

Yan, S., Ren, J., Wang, W., Sun, L., Zhang, W., & Yu, Q. (2023). A Survey of Adversarial Attack and Defense Methods for Malware Classification in Cyber Security. *IEEE Communications Surveys and Tutorials*, *25*(1), 467–496. https://doi.org/10.1109/COMST.2022.3225137

Yang, W., & Yin, F. (2023). A Multi-Strategy Adversarial Attack Method for Deep Learning Based Malware Detectors. *Proceedings - 2023 7th International Conference on Cryptography, Security and Privacy, CSP 2023*, 66–70. https://doi.org/10.1109/CSP58884.2023.00018

Yuan, J., Zhou, S., Lin, L., Wang, F., & Cui, J. (2020). Black-box adversarial attacks against deep learning based malware binaries detection with gan. *Frontiers in Artificial Intelligence and Applications*, *325*, 2536–2542. https://doi.org/10.3233/FAIA200388

Yuan, P., Wang, S., Zhao, C., Wang, W., Bai, D., Peng, L., & Chen, Z. (2023). Adversarial Attack with Genetic Algorithm against IoT Malware Detectors. *IEEE International Conference on Communications*, *2023-May*, 1413–1418. https://doi.org/10.1109/ICC45041.2023.10279299

Zhan, D., Duan, Y., Hu, Y., Yin, L., Pan, Z., & Guo, S. (2023). *AMGmal: Adaptive mask-guided adversarial attack against malware detection with minimal perturbation*. https://www.sciencedirect.com/science/article/pii/S0167404823000135

Zhang, Y., Jiang, J., Yi, C., Li, H., Min, S., Zuo, R., An, Z., & Yu, Y. (2024). A Robust CNN for Malware Classification against Executable Adversarial Attack. *Electronics (Switzerland)*, *13*(5). https://doi.org/10.3390/electronics13050989