

Adversarial Attacks on Artificial Intelligence Models: Defence Strategies, Datasets, Evaluation Metrics, and Future Research Directions

Hassan Mansur Hussien

alabadihassan33@gmail.com

Al-Ameen University

Sharifah Md Yasin

Universiti Putra Malaysia

Nur Izura Udzir

Universiti Putra Malaysia

Mohd Izuan Hafez Ninggal

Universiti Putra Malaysia

Mustafa Thamer Wafeeq

University of Technology-Iraq

Saba I Jawad

University of Bilad Alrafidain

Research Article

Keywords: Adversarial attacks, machine learning, deep learning, cybersecurity, and defence strategies

Posted Date: September 29th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10503793/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

Adversarial Attacks on Artificial Intelligence Models: Defence Strategies, Datasets, Evaluation Metrics, and Future Research Directions

Hassan Mansur Hussien¹, Sharifah Md Yasin², Nur Izura Udzir³, and Mohd Izuan Hafez Ninggal⁴, Mustafa Thamer Wafeeq⁵, and Saba I Jawad⁶

¹Department of Cybersecurity Engineering, College of Engineering, Al-Ameen University, Baghdad, Iraq

^{2, 3, 4}Faculty of Computer Science and Information Technology, Universiti Putra Malaysia, Serdang, Selangor, Malaysia

²Institute for Mathematical Research (INSPEM), Universiti Putra Malaysia, 43400 Serdang, Selangor, Malaysia

⁵Information Technology Centre-University of Technology-Iraq, Baghdad, Iraq

⁶College of Engineering, University of Bilad Alrafidain, Diyala, Iraq

*Corresponding author: alabadihassan33@gmail.com

Abstract

The growing use of machine learning and deep learning models in cybersecurity applications, such as network intrusion detection systems, is remarkable and has a significant impact on the security and privacy of such applications. Therefore, adversarial attacks on these artificial intelligence models can significantly affect malware detection, intrusion prevention, and user authentication by threatening the models through poisoning and evasion attacks. In this survey, the gap between cybersecurity applications and artificial intelligence models was addressed by evaluating adversarial attacks and their defence strategies. A bibliometric analysis of 1,000 research papers from 2017 to 2025 was conducted to identify trends in adversarial attacks on artificial intelligence models. The current survey identified 12 new adversarial attacks on AI models, classified them into categories, and assessed their effects on system security and privacy. Défense strategies to mitigate adversarial attacks on AI models were reviewed and evaluated. This study has discussed key research challenges, such as performance, scalability, defence strategy imbalances, and standardization gaps. Multiple opportunities for future research in terms of technically complex and strategically significant aspects of adversarial attacks on AI models have been highlighted.

Keywords: Adversarial attacks, machine learning, deep learning, cybersecurity, and defence strategies.

List of Abbreviations

Abbreviation	Full Term
AI	Artificial Intelligence
AML	Adversarial Machine Learning
ML	Machine Learning
DL	Deep Learning
DNN	Deep Neural Network
CNN	Convolutional Neural Network
RNN	Recurrent Neural Network
AE	Autoencoder
LSTM	Long Short-Term Memory
DRL	Deep Reinforcement Learning
GAN	Generative Adversarial Network
GNN	Graph Neural Network
XAI	Explainable Artificial Intelligence
IDS	Intrusion Detection System
IoT	Internet of Things
CPS	Cyber-Physical System
DSS	Decision Support System
API	Application Programming Interface
FGSM	Fast Gradient Sign Method
PGD	Projected Gradient Descent
BIM	Basic Iterative Method
C&W	Carlini and Wagner Attack
UAP	Universal Adversarial Perturbation
PCA	Principal Component Analysis
PSO	Particle Swarm Optimization
SVC	Support Vector Classifier
CMAG	Cascade Model-Aware Generative

ADNet	Adversarial Detection Network
APR	Amplitude-Phase Recombination
ER-Classifier	Embedding Regularized Classifier
RSLAD	Robust Soft Label Adversarial Distillation
AMM	Adversarial Margin Maximization
GraphAT	Graph Adversarial Training
PDA	Perturbation Degree Adjustment
GAT	Gradient Alignment Training
DGA	Domain Generation Algorithm
PLC	Programmable Logic Controller
HMI	Human-Machine Interface
VPN	Virtual Private Network
HTTPS	Hypertext Transfer Protocol Secure
TCP	Transmission Control Protocol
UDP	User Datagram Protocol
PCAP	Packet Capture
NLP	Natural Language Processing
AUC	Area Under the Curve
ROC	Receiver Operating Characteristic
FPR	False Positive Rate
FAR	False Acceptance Rate
TPR	True Positive Rate
FNR	False Negative Rate
FRR	False Rejection Rate
TNR	True Negative Rate
FP	False Positive
FN	False Negative
TP	True Positive
TN	True Negative

ASR	Adversarial Attack Success Rate
ODR	Original Detection Rate
EIR	Evasion Increase Rate
EGR	Effective Generation Rate
ER	Evasion Rate
MR	Misclassification Rate
TFR	Target Fooling Rate
AvSR	Average Success Rate

1. Introduction

With the global expansion of digital infrastructure, hackers performed a refined attack on the AI fraud detection system of a bank, stealing \$47 million unnoticed [1]. The mentioned case highlights a rising concern: with the organisations depending on AI for cybersecurity more and more, new threats targeting those AI systems will be faced [2]. Machine learning models now support most of computer security, detecting malware, identifying invaders, and verifying user identities [3]. AI is vital in cybersecurity due to its ability to analyse complex data and adapt to evolving threats. Key AI models (including DNNs, CNNs, RNNs, AEs, LSTMs, DRL, and GANs) are fundamental for automating threat detection, recognising abnormalities, and defending against specific cyberattacks [4], [5], [6], [7], [8], [9]. As shown in **Figure 1**, those models enable scalable and strong, durable cybersecurity through the learning of hierarchical features, detection of time-based patterns, and generation of artificial adversarial samples. Their use in intrusion detection, malware classification, and authentication portrays an innovation in the cyber defence based on AI, with systems detecting up to 95% of threats under normal conditions [10]. But attackers developed adversarial tactics (inputs designed to deceive the AI systems) that are capable of reducing the security effectiveness by 40-80% and making vulnerabilities out of the reliable defences [11].

Adversarial attacks are no longer theoretical only as they can cause real harm. In cybersecurity, these attacks allow malware to bypass detection, poison training data to create backdoors, and expose the AI model's secrets [12]. The attackers use different levels of knowledge: white-box

(full information), black-box (no information), and grey-box (partial information) [13]. Alternatively, adversarial machine learning (AML) inspects how little alterations can sabotage both machine learning (ML) and deep learning (DL) systems during training or conclusion [14], [15], [16], [17]. Techniques like FGSM, PGD, DeepFool, and C&W prove that almost inaudible noise can lead to misclassification, even in the very robust models [18], [19], [20], [21]. Furthermore, adversarial attacks are capable of targeting different AI models. This has created an urgent need to build more robust AI systems, specifically in essential domains such as infrastructure and national security.

The above-mentioned attacks have significantly impacted numerous security domains. For example, computer network applications might lead to the omission of malicious traffic analysis. While the antivirus software relying on the AI model might be incapable of detecting newly discovered malware. One necessary application that might be highly affected is biometric systems, which can be fooled by fake inputs into the AI model [22]. A single incorrect decision by an AI model can lead to a major security failure, particularly in critical systems such as autonomous vehicles and industrial networks [23]. The researcher has been examining AI in the domain of cybersecurity risk, but there is still a need for substantial analysis of adversarial attacks on AI models [24] [25][26][27][28]. The requirement to establish a general framework for evaluating adversarial attacks and their defence techniques in a comprehensive way to balance security, speed and usability remains unaddressed. This limitation has made it difficult for the organisation to choose defensive techniques effectively. Therefore, the current study has developed an essential multi-layered framework by combining technical aspects with bibliometric analysis. The structure of AI model taxonomy has been approached from different aspects of cybersecurity risk, such as attackers in terms of knowledge, timing, and goals. Also, the classification of the efficient defence strategies has been thoroughly examined. The datasets, standards, and evaluation metrics for effectively assessing the defence strategies in AI models against adversarial attacks have been appreciably identified.

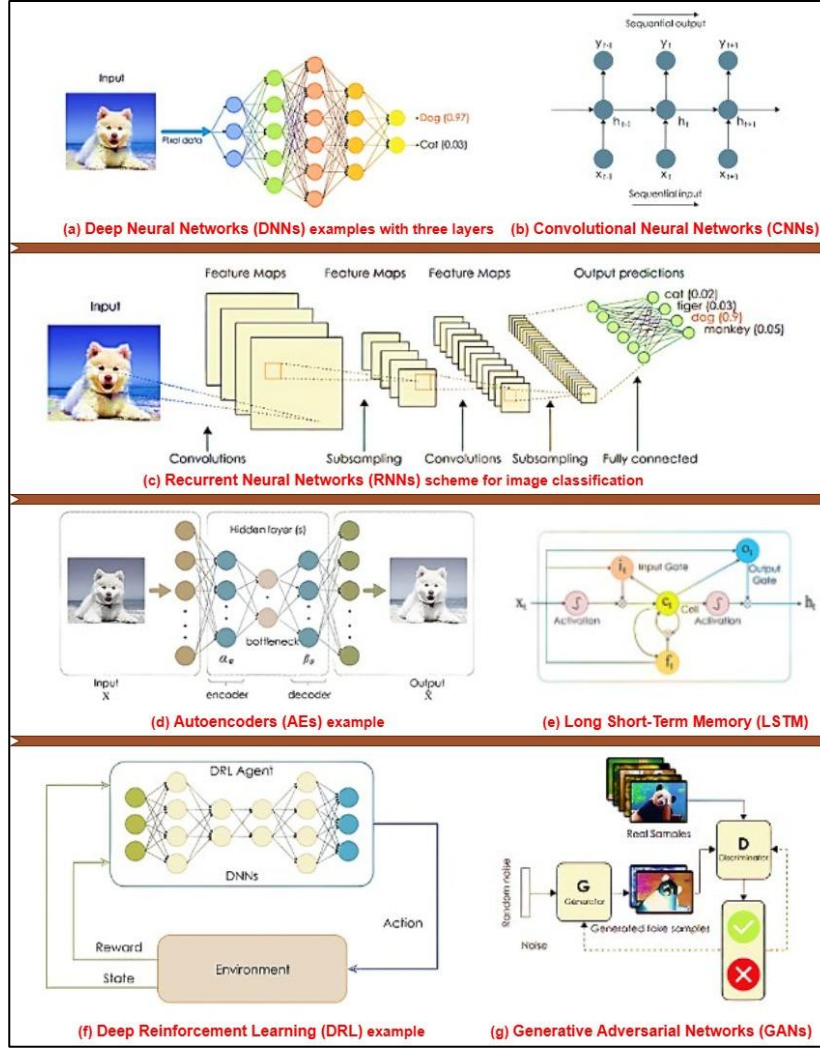


Figure 1 shows the typical structures of DNN, RNN, CNN, AE, LSTM, DRL, and GAN.

The organisation of the paper is as follows. Section 2 covers the related surveys and highlights the main contributions of the current survey. Section 3 presents the research methodology used in this study. The bibliometric overview of the current study is presented in Section 4. The taxonomy of the attacks and the defence strategies have been presented in Section 5. The discussion of the current study is presented in Section 6, and the conclusion is presented in Section 7.

2. Related works and contributions

The investigation of adversarial attacks on AI models has gained significant attention from researchers seeking to examine their robustness. The first framework was proposed in NIST AI 100-2, which offers a general taxonomy of adversarial attacks on machine learning models. This framework provides limited guidelines for the practical deployment of attacks and has failed to examine adversarial attacks and their defence strategies [13]. Machine learning adversarial attacks and their defences have been proposed within the dimensions framework to distinguish between general cybersecurity threats and attacks on AI models, yet an in-depth discussion and technical classification remain lacking [29], [30]. The vulnerabilities and mitigation methods of generative AI in large language models have been examined, but the classification of adversarial attacks remains uncovered [31]. Adversarial examples have been discussed as potential security threats to DL models, yet a comprehensive analysis remains lacking [32]. Adversarial attacks on ML models have been covered, but not with a focus on defence methods against them [33]. **Table 1** tabulated the most recent survey in comparison with the current survey, which has concentrated on the adversarial attacks and their defence methods on the models of ML and DL in several aspects. Based on a critical analysis of previous studies compared with our study, the current study has significantly addressed these gaps by providing a deep-dive analysis of the classification of adversarial attacks and defence methods. The analysis of 1,000 publications from 2017 to 2025 provides a clear picture of the current research trend. The study's framework offers a well-structured taxonomy of adversarial attacks and protection strategies. This tremendous work can serve as a valuable baseline for researchers to build a sufficient and secure AI model system for various practical applications.

Table 1: Comparison of Existing Surveys on Adversarial Attacks

Criteria	Our Paper	[13]	[31]	[32]	[29], [30]	[33]
Focus on Adversarial AI	✓	✓	×	×	✓	×
Cybersecurity Applications	✓	✓	✓	✓	✓	✓
Taxonomy of Attacks	✓	✓	×	×	✓	×
Defense Strategies	✓	✓	×	×	✓	✓
Bibliometric Analysis	✓	×	×	×	×	×
Explainable AI / XAI	✓	×	✓	✓	✓	✓
Federated Learning / Privacy	✓	×	✓	✓	✓	✓

Real-world Deployment Challenges	✓	✓	✓	✓	✓	✓
Standardization / Policy	✓	×	✓	✓	✓	✓
Graph-based Attacks / GNNs	✓	×	×	×	✓	✓
Training vs. Inference Attack Timing	✓	✓	×	×	✓	×
Perturbation Types	✓	✓	×	×	✓	×

3. Research Methodology

This survey has mainly focused on adversarial attacks and their defence strategies for models in neither ML nor DL, or in both. The research methodology has four phases: data extraction, preprocessing, bibliometric analysis, and visualisation, as shown in **Figure 2**. The 1,000-dataset collection has been extracted from 2017 to 2025. These databases have been indexed in Web of Science (WoS) and Scopus. The Bibliometric analysis of the database has been performed by using the open-source R software, specifically the Bibliometrics package. Several studies have been used bibliometric tools to explore fields such as blockchain, healthcare, and cybersecurity, effectively identifying trends and research gaps [34] [35] [36].

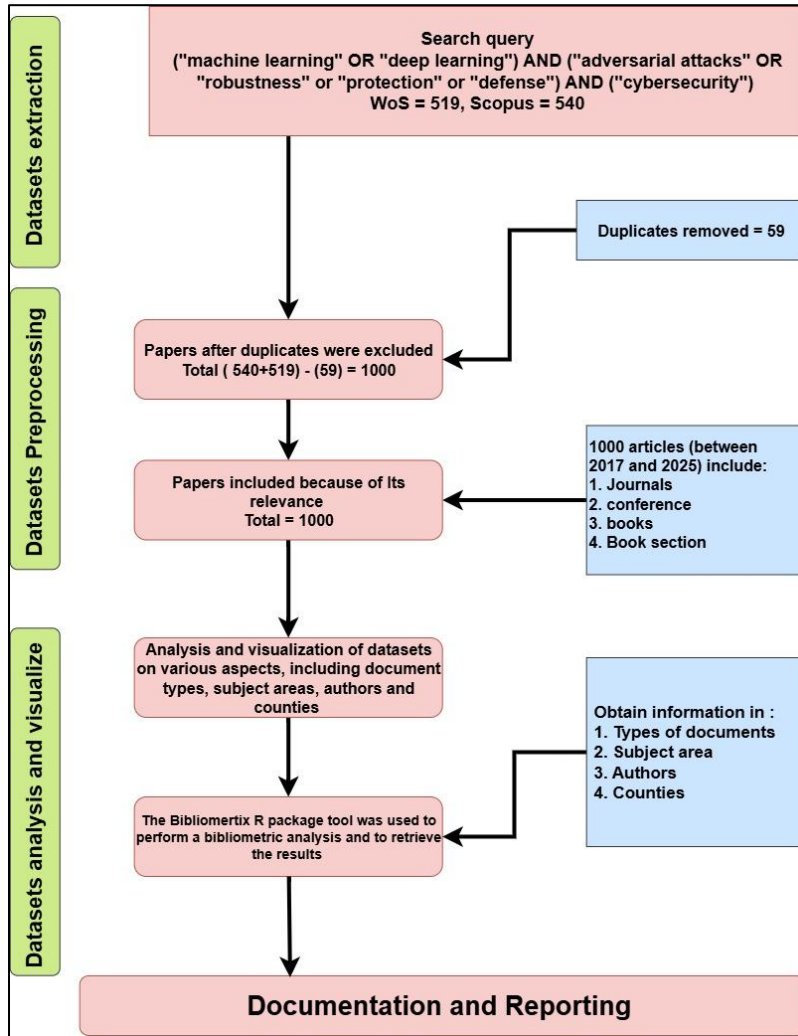


Figure 2: Illustration of the research methodology flow.

4. Bibliometric overview

This section includes a bibliometric analysis of the research field of adversarial attacks and defence in AI within cybersecurity, highlighting main aspects such as: (1) the distribution of datasets, (2) outstanding publication platforms, (3) keyword trends, (4) citation patterns. This analysis aims to help scholars, practitioners, and policymakers understand the development of AI research, recognise major contributions, and therefore make informed decisions regarding future adversarial AI strategies to protect digital systems.

4.1 Datasets distribution

In this study, we analysed a bibliometric dataset that covers the time period from 2017 to 2025, and contains 1000 documents from 563 different publication platforms, including magazines, conference records, book chapters, and other academic forms as shown in Table 2. This analysis showed that adversarial attacks and defences in cybersecurity AI are extensive and involve more fields increasingly. Most of the datasets were published in magazines which indicates increasing academic maturity and precision, which helped in standardising practical research and core ideas. The dynamic and experimental nature of the subject manifests in conference records, and its gaining great importance for mostly including experimental results, and new attack and defence strategies. Additionally, the dataset that includes 30 book chapters, editorials, letters, brief surveys, and 31 review articles highlights the important role of AI in broader cybersecurity discussions. The current study has emphasised the need to examine the AI model in terms of robustness and effectiveness in different applications

Table 2: Main information

Sources	No.
Timespan	2017:2025
Sources (Journals, Books, etc)	563
Documents	1000
Annual Growth Rate %	100.39
Document Average Age	1.41
Average citations per doc	9.314
Keywords Plus (ID)	4155
Author's Keywords (DE)	2175
Authors	3231
Authors of single-authored docs	53
Single-authored docs	76
Co-Authors per Doc	3.87
International co-authorships %	28.5

Figure 3 shows the growth in the integration of the AI model across multiple cybersecurity fields, such as digital forensics, as reflected in publications. However, the AI model is vulnerable to adversarial attacks [37].

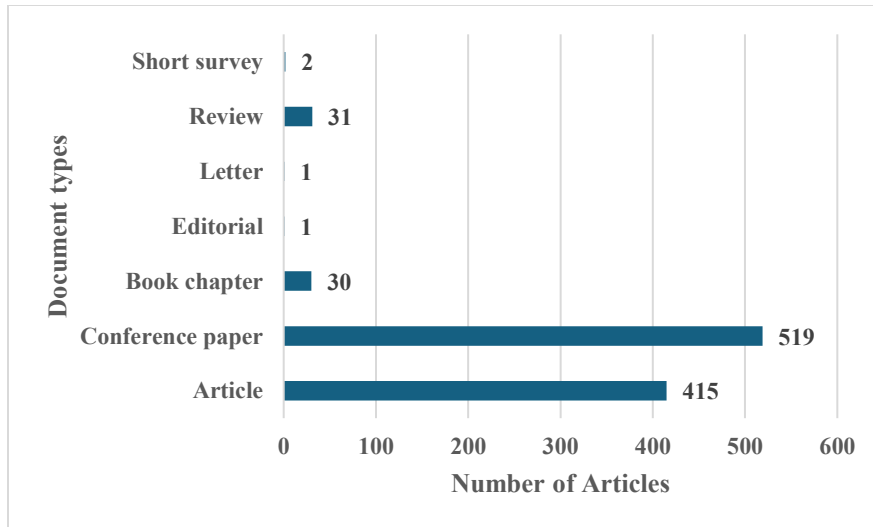


Figure 3: Source classification

Figure 4 shows the annual growth in academic publications on adversarial attacks and their implications. In recent years, the number of publications has steadily risen, with notable peaks indicating significant interest from both industry and academia. As cyber risk has become more complex, the upward trend has recently emphasized the crucial need for secure AI models. The substantial increase in 2019 indicated that AI in field of cybersecurity is rapidly evolving and offering numerous opportunities for security professionals, developers and researchers. **Figure 5** illustrates the annual trend in academic contributions, emphasising the growing significance of adversarial AI for the future of cybersecurity research and practice.

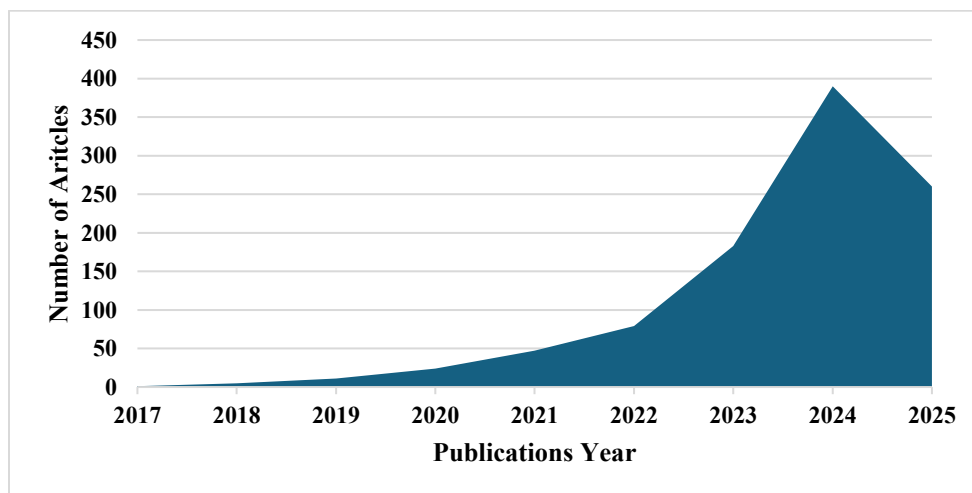


Figure 4: Annual scientific production

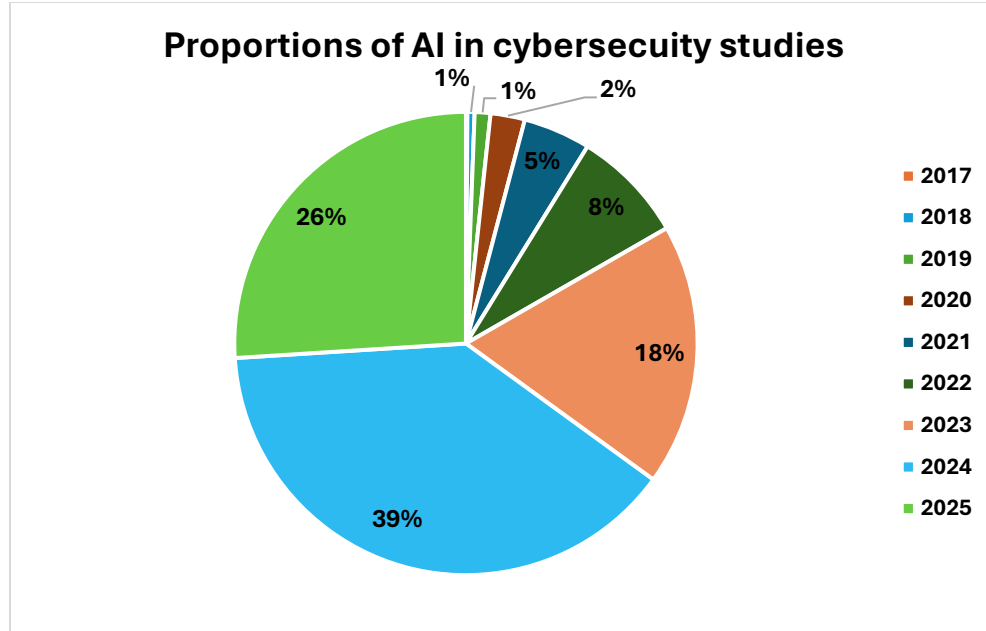


Figure 5: Yearly publication trend

4.2 Top venues

This section categorises academic articles by the journals and conference records that have had a major impact on adversarial attacks and defence in artificial intelligence within the cybersecurity field. This structure helps researchers identify suitable adversarial ML/DL and cybersecurity topics and publication venues. Table 2 lists the top 20 venues in publishing research on adversarial attacks, with IEEE Access leading in volume, followed by Computers and Security, ACM International Conference Proceedings Series, and Lecture Notes in Networks and Systems. Those venues are known for their sustained contributions and play a key role in advancing adversarial AI research. Journals such as the IEEE Internet of Things Journal, Scientific Reports, Sensors, and Expert Systems with Applications have also shown clear engagement with this topic.

Table 3: Top publishing venues that involved adversarial attacks

Sources	Articles
IEEE Access	48
Computers And Security	22
ACM International Conference Proceeding Series	20
Lecture Notes In Networks And Systems	19
Ceur Workshop Proceedings	18
Communications In Computer And Information Science	16

Lecture Notes In Computer Science	14
Applied Sciences (Switzerland)	12
IEE Internet Of Things Journal	12
Scientific Reports	11
Sensors	11
Electronics (Switzerland)	8
Expert Systems with Applications	8
IEEE Transactions on Network and Service Management	8
Computers, Materials and Continua	7
Journal Of Information Security and Applications	7
IEEE Transactions on Dependable and Secure Computing	7
Cluster Computing	5
Computer Networks	5
Future Internet	5

Figure 6 shows the growth of sources over time, displaying an increase in publications, especially within IEEE Access, which is expected to continue expanding its coverage of adversarial AI. The high citation numbers for these venues highlight their impact and the strong link between adversarial machine learning and cybersecurity research.

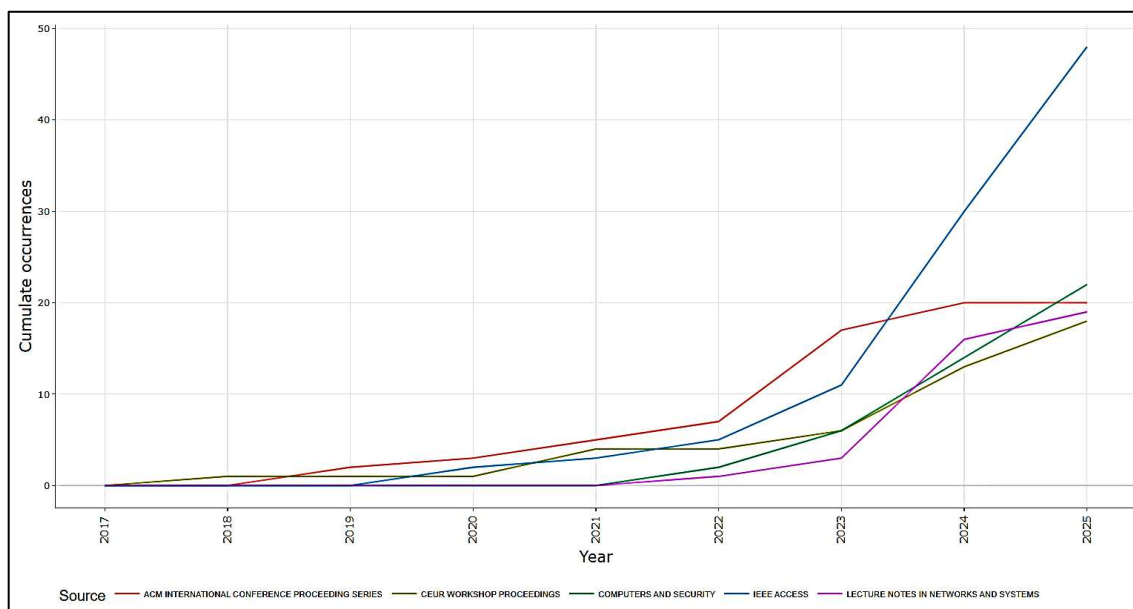


Figure 6: Source growth

4.3 Keyword analysis

The keyword analysis aimed mainly to identify the research gaps in adversarial attacks and their mitigation in AI models. The top 20 keywords that rank most frequently have been listed in Table 4. In this analysis, keywords such as deep learning and attack detection are among the most frequently occurring terms. This indicates that it is vital to develop secure models and defences against adversarial attacks. Additionally, the keywords such as evasion attacks, white-box attacks, and black-box attacks show that researchers emphasise comprehending and countering different adversary strategies. The author has identified a keyword connection between artificial intelligence, distributed systems, and data security, as demonstrated by concepts such as federated learning, secure learning, and privacy-preserving approaches. This suggests that the concept of blockchain is a trending topic worth exploring.

On the other hand, keywords such as GANs, poisoning attacks, model inversion, and differential privacy have been identified as new areas of study and technical challenges. However, the presence of keywords such as intrusion detection systems and malware classification has shown that AI's role in cybersecurity.

Table 4: The most frequently used keywords.

No	Words	Occurrences
1	Cybersecurity	990
2	Deep learning	437
3	Machine-learning	618
4	Adversarial machine learning	280
5	network security	266
6	learning systems	194
7	Intrusion detection	183
8	anomaly detection	124
9	network intrusion	124
10	intrusion detection systems	123
11	computer crime	115
12	cyber-attacks	100
13	intrusion-detection	99
14	generative adversarial networks	97
15	adversarial attack	95
16	malware	92
17	federated learning	83

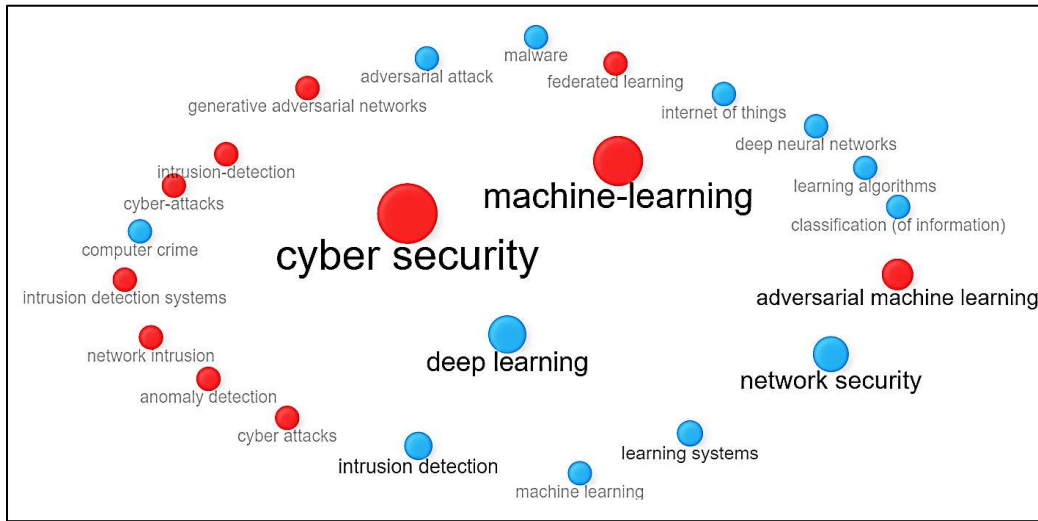


Figure 8: Keyword co-occurrence network.

4.4 Citation distribution analysis

The current section examines the citation distribution of adversarial AI within cybersecurity publications. **Figure 9** shows the top contributing countries by citation count. The United States leads with 801 citations, underscoring its outstanding role in advancing AI research and integrating it into cybersecurity practice. China follows with 693 citations, reflecting the expansion of its investment in AI cybersecurity technologies. Next come India (670), Saudi Arabia (423), and the United Kingdom (413), which also exhibit strong academic engagement through their remarkable contributions to both the theoretical and practical aspects of adversarial learning. The international interest in exploring adversarial attacks on AI models has attracted significant attention, with over 200 citations from New Zealand, France, Australia, and Canada. The exploration spanned different research areas, including secure federated learning, AI-based threat detection, and adversarial robustness and efficiency.

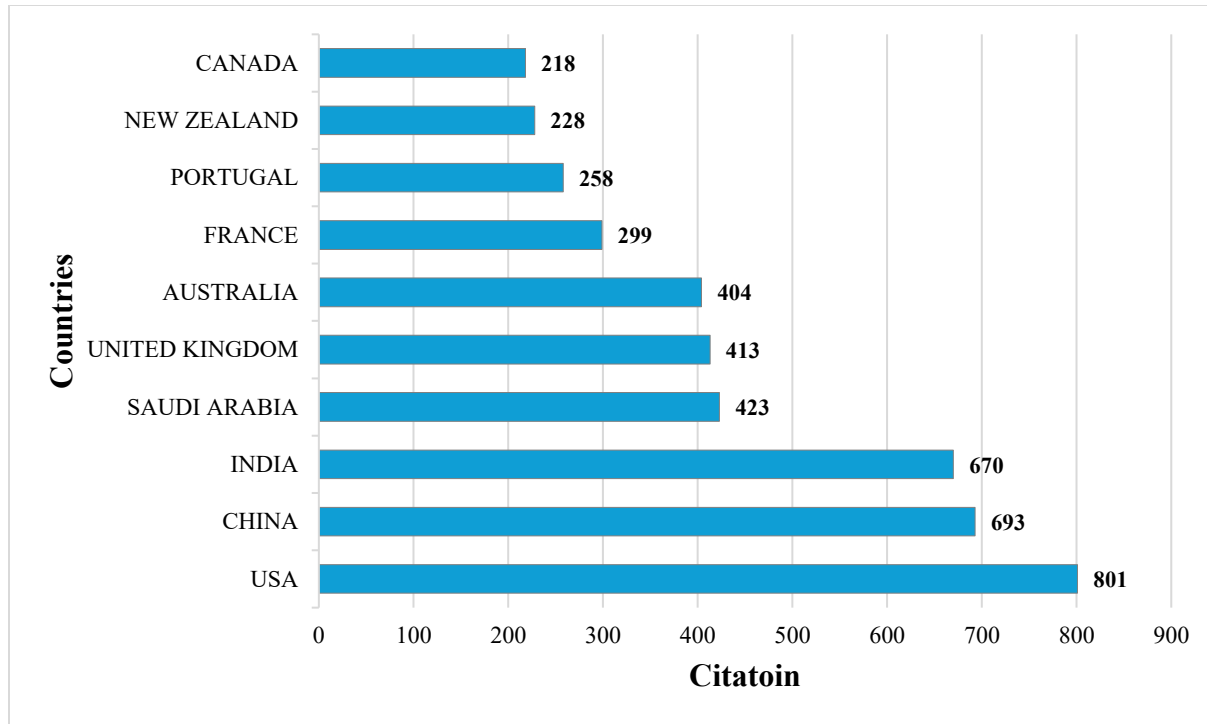


Figure 9: Top countries.

Table 5 tabulates the most cited articles in the area of adversarial attacks on AI models. This approach helps identify the best baseline for this line of research. The highest citation count was 15,867 for a paper proposing a secure adversarial training framework to mitigate white-box attacks on the AI model using Projected Gradient Descent (PGD) [38]. The adversarial attack in deep learning was the second most-cited paper, proposing a secure intrusion detection system for IoT networks in real-time vulnerability traffic classification [39]. Building on the concept of an improved intrusion detection system, a secure defence method was designed to overcome vulnerabilities in distributed IoT networks [40]. Also, a study has been examining the need to build a secure machine learning against adversarial attacks in the case of the cyber physical system [41]. The utilisation of explainable AI (XAI) to enhance the robustness of the AI model was proposed to provide transparency and trust in such a security system [42]. By integrating adversarial robustness into existing security frameworks, a study proposed defence strategies for intrusion detection systems to combat evasion attacks [43]. However, developing a standardised framework to evaluate the adversarial threats in autonomous driving by creating a taxonomy of relevant attacks and defences for industrial informatics has been approached [44]. Also, an interesting study examined machine-generated text and its adversarial challenges, focusing on synthetic content

detection techniques[45]. Adversarial explainability in cybersecurity, AI model interpretability, and robustness were well examined [46]. The discussion of the data poisoning attacks in machine learning and suggested mitigation strategies to protect data integrity during training was highlighted [47].

Table 5: Details of the 10 most cited papers

Ref.	Title	Authors	Publisher	Rank	Year	Citation	country
[38]	Towards deep learning models resistant to adversarial attacks	Madry, A, Makelov, A, Schmidt, L, Tsipras, D and Vladu	6th International Conference on Learning Representations, ICLR	none	2018	15867	USA
[39]	Analyzing Adversarial Attacks against Deep Learning for Intrusion Detection in IoT Networks	Ibitoye, O., Shafiq, O., & Matrawy, A	In <i>2019 IEEE global communications conference (GLOBECOM)</i>	none	2019	330	Canada
[40]	Adversarial attacks against network intrusion detection in IoT systems.	Qiu, H., Dong, T., Zhang, T., Lu, J., Memmi, G., & Qiu, M	<i>IEEE Internet of Things Journal,</i>	IP=9.5	2020	320	China
[41]	Resilient machine learning for networked cyber physical systems: A survey for machine learning security to securing machine learning for CPS.	Olowononi, F. O., Rawat, D. B., & Liu, C.	<i>IEEE Communications Surveys & Tutorials,</i>	IP=46.1	2020	250	USA
[42]	Recent advances in trustworthy explainable artificial intelligence: Status, challenges, and perspectives.	Rawal, A., McCoy, J., Rawat, D. B., Sadler, B. M., & Amant, R. S.	<i>IEEE Transactions on Artificial Intelligence</i>	IP=5.5	2021	230	USA
[43]	Defending network intrusion detection systems	Pawlicki, M., Choraś, M., & Kozik, R.	<i>Future Generation Computer Systems</i>	IP=7.5	2020	210	Poland

	against adversarial evasion attacks.						
[44]	Deep learning-based autonomous driving systems: A survey of attacks and defenses.	Deng, Y., Zhang, T., Lou, G., Zheng, X., Jin, J., & Han, Q. L.	<i>IEEE Transactions on Industrial Informatics</i> ,	IP=9.9	2021	200	Australia
[45]	Machine-generated text: A comprehensive survey of threat models and detection methods.	Crothers, E. N., Japkowicz, N., & Viktor, H. L.	<i>IEEE Access</i> ,	IP=3	2023	192	Canada
[46]	Adversarial XAI methods in cybersecurity.	Kuppa, A., & Le-Khac, N. A.	<i>IEEE transactions on information forensics and security</i>	IP=5	2021	180	Ireland
[47]	Data poisoning attacks against machine learning algorithms.	Yerlikaya, F. A., & Bahtiyar, Ş.	<i>Expert Systems with Applications</i> ,	IP=7.5	2022	118	Turkey

5. Taxonomy of Attack and Defence Strategies in the AI Model

The taxonomy of the current study has been presented in this section. **Figure 10** illustrates the current taxonomy, which consists of a three-tier divide into three groups to enable a clear understanding of adversarial attacks in cybersecurity.

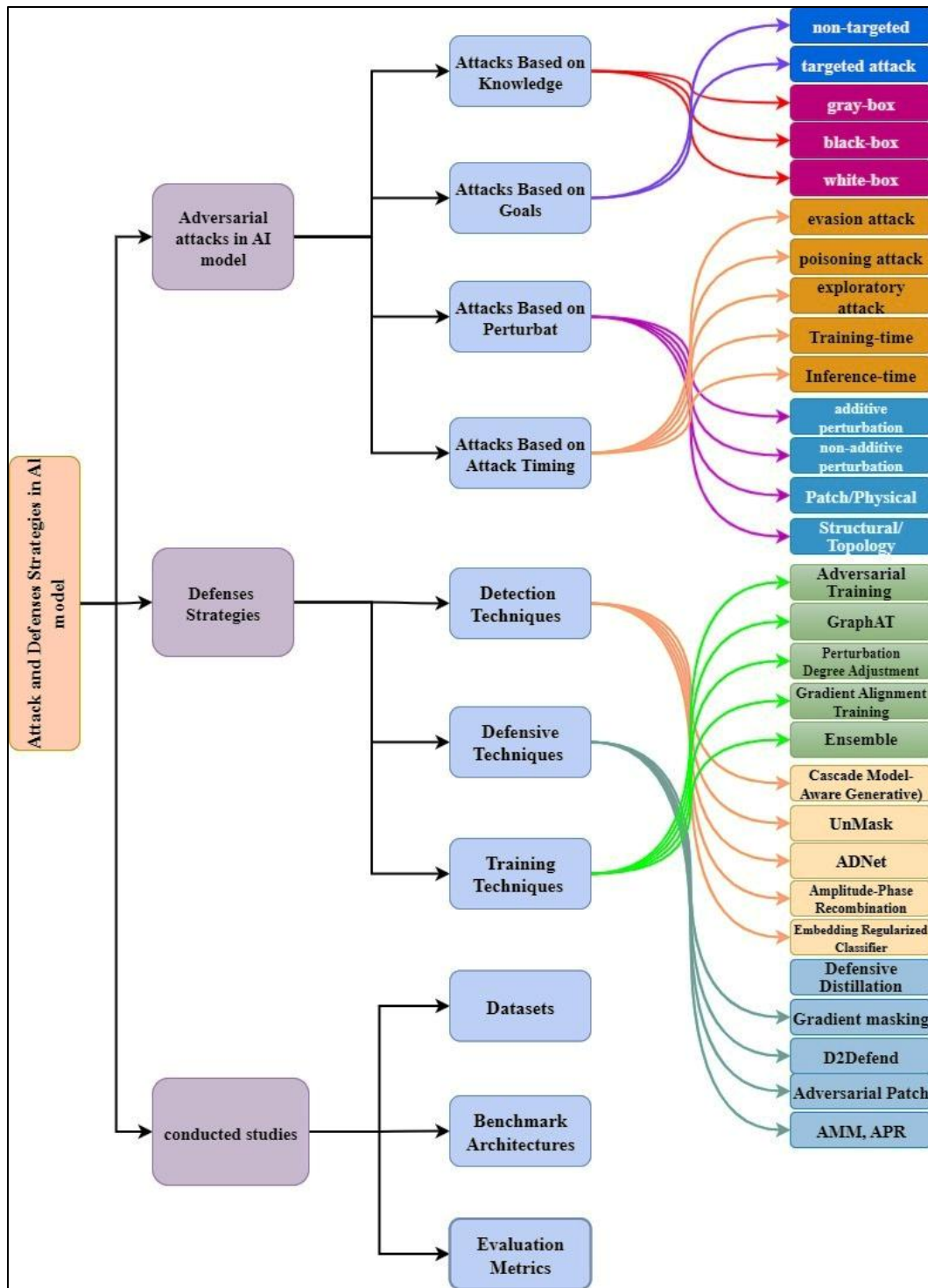


Figure 10: Taxonomy of adversarial attacks and defence strategies

5.1 Adversarial Attacks on AI Models

Adversarial attacks intentionally manipulate input data to fool machine learning (ML) and deep learning (DL) models [48] [49]. These attacks exploit the sensitivity of AI systems to minor changes—often undetectable to humans—that can lead to misclassification or incorrect predictions. In cybersecurity, such attacks threaten systems like malware detectors, intrusion prevention systems, and biometric authentication [50][51][52]. To understand and defend against these threats, adversarial attacks are categorised across four aspects: knowledge, goals, perturbation types, and attack timing. Table 6 offers a comprehensive overview of the adversarial attacks on the modes of AI.

Table 6: Comprehensive review of the literature on Adversarial Attacks in AI Models

Study	Attack Type	Method	Problem Addressed	Advantage	Disadvantage	Major Findings
[53]	White-box	FGSM	Fast generation of adversarial examples using gradient sign	Simple and computationally efficient	Low success rate against robust models	Effective for quick attacks but easily defended against
[54]	White-box	PGD	Iterative improvement over FGSM for stronger attacks	High attack success rate	Computationally expensive	One of the strongest first-order adversaries
[55]	White-box	DeepFool	Minimal perturbation for misclassification	Produces imperceptible perturbations	Less effective on complex models	Efficient in finding close decision boundaries
[56]	White-box	C&W	Optimization-based attack for targeted misclassification	Bypasses many defenses	Slow and complex to implement	Highly effective against defensive distillation
[57]	White-box	LAFeAT	Attacks on latent feature representations	Effective across models	Requires internal model access	Targets internal layers for better transferability
[58]	Black-box	GF-Attack	Graph-based attack without model access	No need for model internals	Requires many queries	Effective on graph neural networks
[59]	Black-box	IPW & IBA	Black-box attacks using iterative queries	Works with limited model access	Query-intensive	Improves attack success with limited queries
[60]	Black-box	NSGA-PSO	Multi-objective optimization for adversarial generation	Flexible and tunable	High computational cost	Balances perturbation and success rate
[61]	Gray-box	SRLIM	Surrogate learning with partial model knowledge	Does not require full model access	Lower precision than white-box attacks	Effective with limited model information
[62]	Gray-box	SCA-based	Exploiting side-channel information	Bypasses traditional defenses	Requires physical or timing access	Can infer model behavior from indirect data
[63]	Gray-box	AtkSE	Error shrinking to mislead models	Subtle and hard to detect	Slow convergence	Reduces model confidence gradually
[64]	Targeted	C&W	Optimized misclassification to a specific target class	Precise control over output class	Computationally expensive	Highly effective in bypassing defenses and achieving targeted misclassification

[65]	confidence confidence	AdvPatch	Physical-world targeted misclassification using adversarial patches	Works in physical environments	Requires physical access or deployment	Effective in fooling models in real-world scenarios
[66]	Targeted	AdGAN	Mislead reinforcement learning agents toward specific wrong actions	Applicable to sequential decision-making systems	Complex to implement and train	Demonstrates vulnerability of RL agents to targeted adversarial manipulation
[67]	Untargeted	FGSM	Cause misclassification with minimal perturbation	Computationally efficient	Less effective against robust models	Fast and simple method for generating adversarial examples
[68]	privacy invasion	BIM	Iterative misclassification using small perturbations	Higher success rate than FGSM	Slower due to iterative nature	Improves attack success over FGSM
[69]	Denial of service	PGD	Robust untargeted attack using projected gradients	Strong attack performance	Requires tuning of parameters	Widely used benchmark for adversarial robustness
[70] [71]	Untargeted	DeepFool	Minimal perturbation for misclassification	Low perturbation magnitude	Less effective on complex models	Efficient in finding decision boundaries
[72]	Untargeted	UAP	Universal perturbation applicable across inputs	Reusable across samples	Lower success rate than input-specific attacks	Perturbation generalizes across multiple inputs
[73]	model extraction	FeaCP	Gradient-free attack aiming for boundary crossing	Works in black-box settings	Requires heuristic tuning	Effective without gradient information
[74] [75] [76] [77] [78]	additive perturbation	FGSM, PGD, DeepFool, LAFeAT, DSNGD	Misclassification through imperceptible noise	Simple implementation; high success rate in gradient-access scenarios	Limited transferability; vulnerable to adversarial training defenses	Effective in white-box settings; fast generation of adversarial examples
[78][79] [79]	Patch/Physical	Adversarial Patch, AdvCam, MAP, RP2	Real-world attacks using visible modifications	Applicable in real-world scenarios; does not require model access	Visibly detectable; limited stealth in digital-only contexts	Can fool models in physical environments; robust to transformations
[80] [81]	Feature/Latent	SSAH, AoA	Manipulation of internal feature representations	Improved transferability; harder to detect with input-level defenses	Requires deeper model understanding; complex implementation	Targets high-frequency and attention components for stealthy attacks
[82] [83] [84]	Structural/Topology	GF-Attack, IGA, SGA	Graph-based attacks altering node/edge structures	Applicable to structured data; bypasses input-level defenses	Limited to graph models; requires graph-specific knowledge	Effective in disrupting graph neural networks; edge flipping strategies
[85], [86]	poisoning attack	AdMRL	Poisoning reinforcement learning agents during training	Effective in disrupting long-term learning behavior	Requires access to training pipeline and environment	Adversarially trained agents can be misled to learn harmful policies
[86]	Training-time	Graph Node Injection	Injecting malicious nodes into graph structures during training	Targets structural vulnerabilities in GNNs	Limited to graph-based models	Can degrade performance of graph neural networks
[38] [87] [88]	evasion attack	FGSM, BIM, PGD, DeepFool	Evasion attacks during model inference	Applicable to deployed models without retraining	Less effective against adversarially trained models	Models can be misled by imperceptible perturbations at test time
[89]	exploratory attack	Physical Attacks (patches, laser beam)	Manipulating sensor inputs in real-world environments	Effective in physical deployment scenarios	Visually detectable and environment-dependent	Can cause misclassification in autonomous systems

[90]	Inference-time	Text/Voice Hidden Command Attacks	Embedding hidden commands in benign inputs	Stealthy and hard to detect	Requires precise encoding and context-specific tuning	Can trigger unintended actions in voice/text-based systems
------	----------------	-----------------------------------	--	-----------------------------	---	--

Figure 11 illustrates how various factors influence adversarial attacks on AI models. These factors include the attacker's knowledge of the model (white-box, black-box, grey-box), the timing of the attack (training-time poisoning versus inference-time evasion), the type of perturbation (additive noise, physical patches, latent feature manipulation), and the attack's objectives (targeted misclassification versus untargeted disruption) [50][51][52].

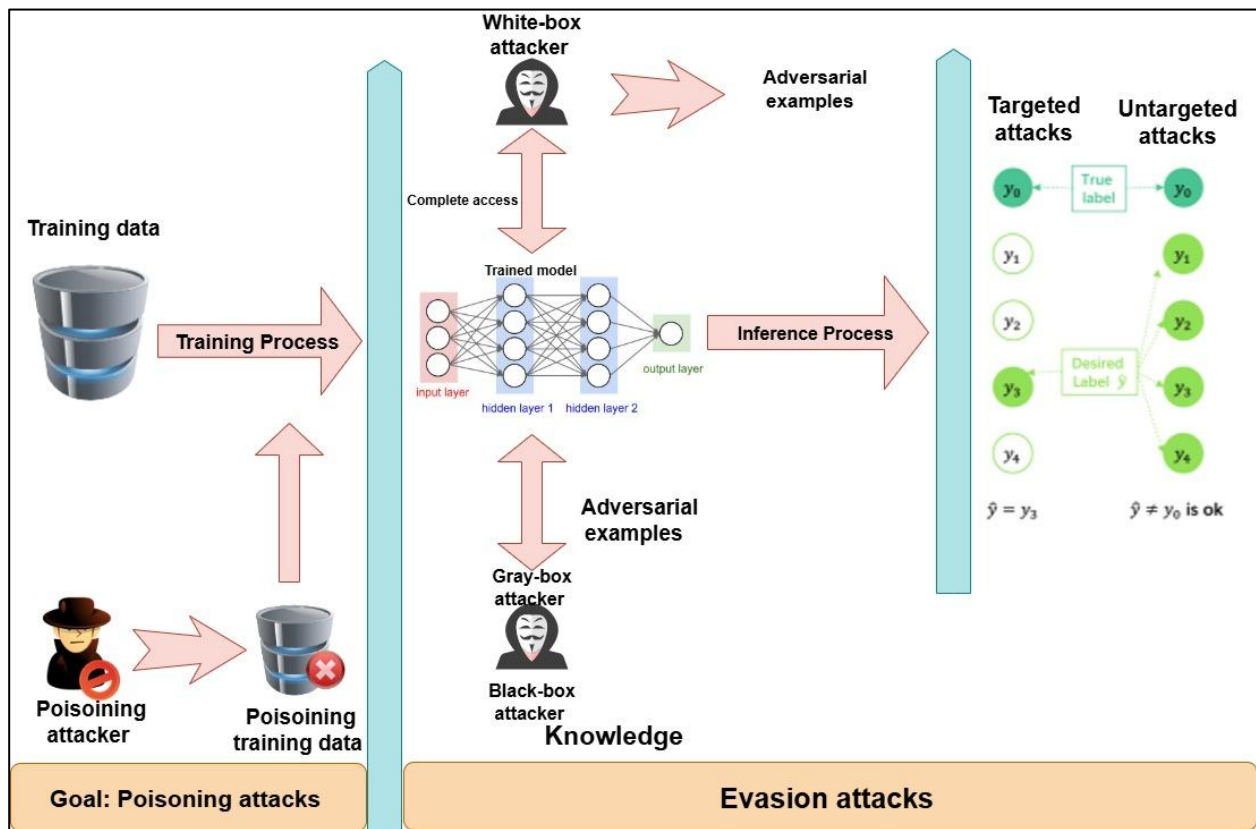


Figure 11: Illustration of adversarial attacks on AI models.

5.1.1 Attacks Based on Knowledge

White-box attacks: This type of attack has the ability to access the complete knowledge of the AI model in terms of structure, parameters, and gradients. For example, GSM, PGD, DeepFool, C&W attack, and LAFEAT method are highly effective and primarily can target systems such as malware classifiers, biometric authentication, and intrusion detection [53] [54] [55] [56] [57]. In fact, these methods exploit the weakness found in the internal model, but they require high

processing power and can be less feasible in practical scenarios. However, preventing this sort of attack is essential because it might pose a significant cybersecurity risk, especially in autonomous vehicles and industrial control systems.

Black-box attacks: This type of attack does not rely on knowledge of the model's internal workings. Instead, they deduce behaviour through input-output interactions. Techniques such as GF-Attack, IPW & IBA, as well as NSGA-PSO [58] [59] [60] generate adversarial examples without relying on gradients. This form of the attack has the ability to significantly expose the application of encrypted traffic analysis, a fraud detection system that relies on AI model training. This attack can interface with the AI model through its application APIs. In matters of fact, they are particularly less concerned than black-box attacks about their impact on real-world applications, but it remains important to prevent them.

Gray-box attacks: This type of attack can expose the AI model by accessing either the training data, the architecture, or both, without referring to the model's parameters. The adversarial attack examples embedded in the datasets used for model training are generated using AtkSE, SCA, and SRLIM [61] [62] [63]. This sort of attack can expose the IoT network, access control system, and federated learning that rely on AI models; as a result, semi-disclosure occurs frequently. The sharing of the AI system in the semi-transparent field is found to be compromised in the Gray-box attacks due to their subtlety, which makes the AI model detection being inaccurate

5.1.2 Attacks Based on Goals

Targeted attacks: This attack works by tricking ML and DL models into misclassifying dataset inputs, which could lead to serious security gaps in such systems. The real-world scenarios that could be affected by these types of attacks include impersonation and bypassing verification, which could compromise users' authentication, biometric recognition, and autonomous decision-making through malicious misclassification, leading to identity theft or unauthorised access. The way of generating adversarial examples is by using methods such as Carlini & Wagner (C&W), [64]AdvPatch [65], and [66] AdGANs are later used to train the AI model to cause misclassification in the dataset.

Untargeted attacks: This type of attack compromises the AI model by targeting a specific class without access to the entire dataset. They are commonly used against intrusion detection systems (IDS), malware classifiers, and network traffic analysers, where creating general disruption helps evade detection. Such techniques include FGSM, BIM, PGD, DeepFool, UAP, and FeaCP, which aim to broadly impair model performance within the AI model [67][68][69][70] [71] [72] [73]. For example, in IDS scenarios, untargeted attacks can cause malicious traffic to be misclassified as benign, allowing threats to bypass detection. These attacks are frequently utilised in adversarial scenarios in the real world because they are simpler to execute and can be scaled more easily.

5.1.3 Attacks Based on Perturbation Types

Additive perturbation attack: This type of adversarial attack on DL and ML models works by adding imperceptible noise to the dataset. This form of attack is used in white-box scenarios when it tricks the system, such as an IDS, into detecting the malware, causing misclassification of the files. These methods are FGSM, PGD, DeepFool, LAFeAT, and DSNGD, and yet they require computational resources to launch the attacks [74] [75][76][77][78]. .

Patch perturbation attacks: This type of attack works by making visible changes to the dataset of the AI model by using these adversarial examples such as Adversarial Patch, AdvCam, MAP, and RP2 [91][79]. This form of attack can directly affect real world scenarios, such as biometric systems, autonomous vehicles, and surveillance systems.

Feature-perturbation-level attack: This adversarial attack on ML and DL models works by perturbing the models' internal representations. This attack aimed to disrupt the attention maps or latent features of the models. The methods used are SSAH and AoA to fool the model by producing high frequency in the semantic elements [80][81]. Mitigating this sort of attack is considered a challenge to detect because it can transfer across different models easily. The attack can significantly impact identity verification and access control by compromising the integrity of the system's internal features.

Structure-based perturbations: This type of attack could target the AI model by manipulating the node and edge relationships in the data model's structure graph. The methods for generating

adversarial examples are GF-Attack, IGA, and SGA to manipulate graph neural networks (GNNs). This sort of attack mainly affects the application of botnet detection, social network security, and network traffic analysis [81] [83] [83].

5.1.4 Attacks Based on Attack Timing

Poisoning Attacks: This is a type of adversarial attack that works on ML and DL models by modifying the training data. This attack works by injecting malicious data into the original dataset, which could lead to massive vulnerability in the model by using one of these methods AdMRL [85], [86] and Graph Node Injection [86] . In cybersecurity, such attacks affect network intrusion detection systems, IoT security infrastructures, and cyber-physical systems, where poisoned data can lead to misclassification, backdoors, or gradual performance degradation.

Evasion Attacks: This form of adversarial attack can work only after the model has already been trained. It can be launched by crafting adversarial input and fooling the model during the prediction stage. The methods used to generate these adversarial examples are FGSM, BIM, PGD, and DeepFool [38] [87] [88]. These forms of attacks can cause security breaches or impersonation of the system in applications of biometric authentication systems and malware detection systems that rely on AI model training,

5.2 Defence Strategies in AI Models

AI defence strategies are categorised into three primary types to counter adversarial threats in cybersecurity: detection techniques, defensive techniques, and training techniques as shown in Table 7.

Table 7: Current research on AI model defence Strategies

Reference	Method	Problem Addressed	Major Findings	Advantage	Disadvantage
[92]	CMAG (Cascade Model-Aware Generative)	Detect adversarial inputs before model inference	detects adversarial perturbations using logit-based features	Early detection and filtering of adversarial samples	May require additional computation or sensor data
[93]	UnMask	embedded evasion attacks before model inference	reconstructs clean images to detect adversarial noise	Early detection and filtering of embedded evasion attacks before model inference	Extra computation requires

[94]	ADNet	Missclassification of detect adversarial dataset	an anomaly detection network for adversarial samples	Detection and filtering of Missclassification	Require Extra computation in limited resources setting
[95]	APR (Amplitude-Phase Recombination)	Detect poisoning adversarial during training	filters high-frequency adversarial noise	Early detection of poisoning adversarial during training	Additional computation is needed in sensor
[96]	ER-Classifier (Embedding Regularized Classifier)	Detect feature manipulation of adversarial during training of model	detects unusual embedding features	Detection of feature manipulation of adversarial during training of model	Insufficient for sensor data due extra computation is needed
[97], [98]	Defensive Distillation	Gradient-based adversarial attacks	Smooths model gradients to reduce sensitivity to perturbations	Improves robustness without retraining	Limited effectiveness against strong attacks
[99], [100]	Gradient Masking	Gradient exploitation by attackers	Obscures gradient information to hinder attack generation	Simple to implement	Can lead to false sense of security
[101]	D2Defend	Vulnerability in both input and latent spaces	Dual-domain defense combining input and latent space protection	Comprehensive coverage	Increased computational complexity
[102]	RSLAD	Model sensitivity to adversarial perturbations	Uses soft label distillation to enhance robustness	Improves generalization	Requires careful tuning
[103]	AMM	Narrow decision boundaries vulnerable to attacks	Maximizes adversarial margin to strengthen decision boundaries	Enhances robustness against multiple attacks	May affect model accuracy
[104], [105]	Adversarial Patch Robustness Training	Physical-world adversarial patches	Trains models to recognize and resist patch-based attacks	Improves physical robustness	Requires extensive data augmentation
[106], [107], [108]	Adversarial Training (FGSM, PGD)	Model vulnerability to gradient-based adversarial attacks	Improves robustness by retraining with adversarial examples	Widely applicable and effective against known attacks	Computationally expensive and may overfit to specific attack types
[109], [110]	GraphAT	Adversarial robustness in graph neural networks	Enhances node classification accuracy under attack	Tailored for graph-based data	Limited to graph domains
[111]	PDA (Perturbation Degree Adjustment)	Overfitting to strong perturbations during adversarial training	Balances perturbation strength to improve generalization	Improves training stability	Requires careful tuning of perturbation parameters
[112], [113], [114]	GAT (Gradient Alignment Training)	Gradient misalignment between clean and adversarial examples	Aligns gradients to improve robustness	Enhances model interpretability and robustness	May reduce performance on clean data
[115], [116], [117], [118], [119]	Ensemble	Single-model vulnerability to diverse attacks	Combines multiple models to improve robustness	Improves generalization and attack resistance	Increased computational and memory cost

5.2.1 Detection Techniques

This technique is used to improve the security of AI models against adversarial attacks. At the same time, AI models are used to develop intrusion detection systems (IDS), malware classification tools, and biometric authentication platforms to ensure the security of applications.

Therefore, the development of a secure AI mode is crucial for enhancing decision accuracy and ensuring system integrity.

UnMask: This is one of the techniques used to detect adversarial attacks in machine learning models. It works by restoring potentially adversarial example inputs to their clean data during training to reveal hidden disruptions in datasets. UnMask aims to prevent unauthorised users from accessing systems, like biometric authentication and surveillance. This method can ensure the results are protected from spoofing, and evasion attacks. In scenarios of image detection, the UnMask attempts to rebuild datasets of images by comparing them to the originals and identifying adversarial examples to detect abnormality. These procedures can enhance accuracy and preserve the integrity of the data visualisation [93].

Cascade Model-Aware Generative (CMAG): This is a novel detection technique that effectively differentiates between the adversarial examples and the clear dataset. This method works by making predictions and analysing logit properties to identify adversarial inputs in an artificial intelligence model. It uses the autoencoder as a generative model to ensure robust machine learning. This method is suitable for the development of malware detection and intrusion prevention systems that rely on machine learning, whereby attackers attempt to bypass classification. CMAG is not qualified for a complicated dataset due to the increased processing power required [92].

Adversarial Detection Network (ADNet): This is one of the detection techniques used to distinguish between adversarial samples and clean samples in irregularity detection. The ADNet method works by detecting adversarial attacks in irregular detection based on a weighted conditional gradual adversarial network (WCSAN) with a particle swarm optimisation (PSO) algorithm and support vector classifier (SVC) for classification. The feature selection and extraction used the principal component analysis (PCA) algorithm to optimise the parameters of the generator and discriminator. This method can improve the adversarial training of the system. The development of anomaly detection relies on AdNet, which has the ability to improve systems against emerging threats by enabling active protection in dynamic applications [94].

Embedding Regularised Classifier (ER-Classifier) : This is a detection technique that improves the classification of dataset accuracy against adversarial examples. ER-Classifier works by embedding adversarial examples using a regularisation AI model. The extracted feature is regularised by using a discriminator in the latent space, which attempts to distinguish between the generated code vectors and the ideal code vectors. This method is suitable for cybersecurity applications such as malware detection, biometric verification, and phishing defence that involve text, images, or audio. In the scenarios, the system can observe variations from normal embedding patterns by flagging suspicious data prior to completing predictions. The development of the ER-Classifier in such a system will be able to improve the model's strength and facilitate early threat detection. Despite the fact that this method may demand auxiliary processing, implementing it in a cybersecurity application can definitely improve protection against malicious attacks [96].

Amplitude-Phase Recombination (APR): This is a detection method that aims to construct effective training examples based on a few pairs of samples, thereby increasing security against adversarial attacks. APR tends to assemble at the local optimum, which is nearly connected to the high-frequency elements of the training images. At the same time, the amplitude spectrum is efficiently disturbed by noise or typical corruptions. This method involves making remarkable observations to guarantee the generalised behaviours in both the sample adversarial dataset and the clear dataset by recombining the phase spectrum of the current image with the amplitude spectrum of the distractor image. This method is suitable for ensuring the security of audio cybersecurity applications, whereby attackers target specific signal frequencies, such as those used in voice authentication and spoofing detection. The development of the APR decreases noise whilst maintaining authorised communications by recombining phase and amplitude components to improve detection accuracy and system robustness [95].

Table 7 reviews the researcher papers on detection techniques and delivers a systematic comparison of recent methods that aim to strengthen AI security against evasion attacks. These techniques aim to identify and eliminate adversarial perturbations that could impact AI model results. UnMask works perfectly with visual data by reconstructing the clear images to expose hidden noise. CMAG aims to identify adversarial patterns transparently by using logit-based feature selection. ADNet improves anomaly detection by differentiating between adversarial examples and the clear dataset. APR prevents high-frequency attacks by filtering the dataset and

distinguishing between malicious and benign data. ER-Classifier enhances robustness at the feature level to improve the detection of anomalies in the embedding space. In conclusion, all these detection methods are essential in the development of a secure AI model in the cybersecurity application by facilitating early detection.

5.2.2 Defensive Techniques

Defensive techniques can ensure the security of the artificial intelligence model against adversarial attacks. These sorts of proposed solutions are very important in the application of malware classification and intrusion detection systems

Gradient masking: This is a simple defence technique that creates adversarial examples to protect the hidden gradient data of an AI model against adversarial attacks. However, these methods have drawbacks in terms of accuracy, which may introduce a false positive in terms of security detection, and slightly can still lead to the attacks succeeding. This gap can lead to attacks to interfere with the optimisation of AI models such as PGD and C&W. This method is suitable for cybersecurity applications such as biometric identification and fraud detection. This is because it is easily implemented and tremendously reduces the relevance of the model's gradients. The development of AI models relies on gradient masking methods that provide a balance between security and performance. Thereby, it is making them more suitable for limited resource-constrained environments, particularly when reducing computational demands is paramount [99], [100].

Defensive Distillation : This is an innovative strategy for the detection and mitigation of adversarial examples to deceive the model at the time of testing. This method works by combining input sanitisation and ensemble methods to improve the model's ability to classify inputs accurately even in the presence of adversarial perturbations. This method enables the security of AI models against gradient attacks during dataset training, whereby threats such as FGSM and PGD are involved. This method is suitable in cybersecurity domains, such as intrusion detection and malware classification, because the AI model does not need to be retrained. Despite the fact that the development of this method into such a system may be less effective against more sophisticated threats. Therefore, this method remains an important solution for defensive AI

models in cases where attackers manipulate system vulnerabilities to bypass detection or impersonate authorised users [97], [98].

D2Defend : This is a dual-domain-based defence method that aims to recover low-frequency and high-frequency image structures in both spatial and transform domains. This method works by removing the adversarial perturbations by decomposing the input image into edge feature and texture feature layers. Additionally, the method in this study uses two-sided filtering and short-time Fourier transform to displays its effectiveness against multiple adversarial attacks. it is suitable for analysing encrypted traffic and spotting cybersecurity abnormalities because adversarial attacks usually target inner activations or raw data. The development of the D2Defend into an AI model is important for ensuring the robustness of such a system, but it is a significant computational process to accomplish the training [101].

Robust Soft Label Adversarial Distillation (RSLAD) : This is a novel adversarial robustness method that trains AI models from small datasets. This method works by exploiting the robust soft labels produced by an adversarial trained model to differentiate between adversarial and clean on such a dataset. The method has been empirically demonstrated to improve the robustness of small models against adversarial attacks. RSLAD is suitable for cybersecurity fields, such as identity verification and access control, where accuracy and reliability are crucial. The development of RSLAD in AI mode is an important part of adversarial defence because it can protect against both white-box and black-box attacks [102].

Adversarial margin maximisation (AMM) : This is a defence technique that aims to ensure the security of an AI model against DeepFool adversarial disturbance by using a proxy regularisation. AMM works by investigating the classification margin in the input dataset and deriving a novel and principled regularise to expand it. The classification has been developed by using aggregation functions and shrinkage functions to improve the classification margin directly. AMM is suitable for cybersecurity applications, such as autonomous systems and industrial control networks, to ensure accurate decision-making in these systems. The development of the AI model on AMM is critical because it can handle threats that use gradients and patches, but it suffers from slow performance on clean data [103].

Adversarial Patch Robustness: This is one of the defensive techniques that aims to defend the model against physical adversarial patches. This method works by combining the adversarial patches into the training data of AI in order to improve their resistance to detection evasion. Adversarial Patch resistance is crucial in the era of cybersecurity applications, including surveillance, facial recognition, and smart access control, as it enables the recognition and mitigation of malicious visual patterns that could deceive camera systems [104],[105].

Table 7 provides a summary of research papers on defensive techniques. These methods aim to enhance the artificial intelligence model to be more secure against adversarial attacks. The main idea of these strategies is to modify or construct a model to reduce its vulnerability to attack.

5.2.3 Training Techniques

Training defence techniques are able to ensure the resilience of the AI model against adversarial attacks. They are frequently employed in cybersecurity to detect fraud, classify malware, and identify intrusions, where it is essential to counter new attack methods.

Adversarial Training : This is the most common approach to improve a model's robustness against adversarial attacks. This method employs adversarial examples generated by methods like FGSM and PGD to train models on the clean dataset and adversarial dataset. Adversarial training is particularly beneficial in cybersecurity for malware classifiers and intrusion detection systems, where continuous operation during an attack is crucial. It remains a key tactic for developing robust AI systems in challenging environments, despite being computationally intensive and sometimes leading to overfitting on specific types of attacks [106], [107], [108].

Graph Adversarial Training (GraphAT) : This is one of the Training defence techniques aimed to increase the robustness of graph neural networks through the use of perturbations to the graph structure during training (GNNs). This technique enhances node classification accuracy in adversarial environments, which is highly beneficial for various cybersecurity domains, including botnet detection, network traffic analysis, and social network security. The development of the GraphAT in an AI model can ensure the robustness of the system by introducing adversarial noise to nodes and edges. [109], [110].

Perturbation Degree Adjustment (PDA) : This is a novel adversarial training method that aims to ensure the security of AI against adversarial attacks, particularly those from the FGSM adversary. This method aims to learn a representation of semantically meaningful domain invariance on both the clean and adversarial datasets. This method is suitable for applications such as encrypted traffic analysis and fraud detection. This is because it can balance the accuracy and robustness of the system. The development of the PDA allows models to handle various attack techniques without compromising performance on normal data [111].

Gradient Alignment Training (GAT) : This is a training technique that aims to enhance the robustness of DNN models in semantic segmentation by combining adversarial training with domain generalisation theory. This method works by incorporating a domain-adaptive mechanism that dynamically modulates input data, effectively reducing adversarial disturbance. This method is suitable for cybersecurity fields such as identity verification, access control, and abnormality detection. This is because it improves segmentation accuracy on clean images and extremely improves robustness against adversarial attacks [112], [113], [114].

Ensemble training : This is a technique method aimed at enhancing the artificial intelligence model's resilience against a variety of adversarial scenarios. This method works by employing various architectures of AI to train the dataset, thereby improving the robustness of the model. Ensemble training is essential in cybersecurity systems, such as autonomous vehicles and industrial control networks. The development of the system relies on the ensemble training strategy, which can effectively protect the system against known or unknown threats [115], [116], [117], [118], [119].

Table 7 summarises the research papers on the training of defence techniques. These techniques can ensure the resilience of AI models against both white-box and black-box attacks. They are essential in adversarial machine learning. Because they assist the development of AI systems that are trustworthy and robust.

5.3 Conducted Studies

Datasets, standard architectures, and evaluation metrics are presented in this section and they are important to evaluate the adversarial attacks and defences.

5.3.1 Datasets

The most common datasets that are used to evaluate models' resilience are identified in this section. Techniques such as deep learning and machine learning are used to counter adversarial attacks. These datasets have been used to test the AI model in various cybersecurity areas, including user identification, user authentication, intrusion detection, cyber-physical systems, fraud detection, encrypted traffic analysis, and malware detection. Therefore, employing these datasets on models is imperative to demonstrate the ability of new adversarial techniques. Table 8 outlines the primary datasets used for developing and accessing adversarial machine learning-based deep learning security systems, grouped by application.

Table 8: Common datasets for cybersecurity

Dataset	Application	Description
Androzoo [120]	Malware detection	This dataset contains over 16 million apps collected from various sources, such as the Google Play app market.
IoTPOT [121]	Malware detection	This is a dataset with 500 IoT malware samples extracted from an IoT honeypot.
Genome [122]	Android Malware	This dataset has around 1242 malicious Android applications divided into 49 malware families.
Drebin[123]	Android Malware	This dataset contains Android apps from 179 distinct malware families.
DGArchive [124]	Botnet detection	This dataset enables the resolution or calculation of domain names that malware dynamically creates using DGAs.
UMUDGA [125]	Botnet detection	This dataset includes more than 30 million domain names that have been manually labeled and generated through algorithms.
BoT-IoT [126]	Intrusion Detection system	This dataset contains benign IoT-related traffic along with other network traffic, as well as different types of malicious traffic.
CIC-IDS2017 [127]	Intrusion Detection system	This dataset includes labeled network flows with complete packet payloads in PCAP format, along with the associated profiles and CSV files.
CIC-DDoS2019 [128]	Intrusion Detection system	This dataset includes various DDoS attacks conducted through application layer protocols using TCP/UDP.
CSE-CIC-IDS2018 [129]	Intrusion Detection system	This dataset contains records of the following attack types: Brute-force, Heartbleed, Botnet, DoS, DDoS, Web attacks, and internal network infiltration.
MedBIoT [130]	Intrusion Detection system	This dataset was assembled from a medium-sized network of 83 devices, including both real and emulated IoT devices.
BATADAL[131]	Cyber-Physical System	This dataset contains 8761 records covering 43 variables related to a water distribution network.
Gas pipeline [132]	Cyber-Physical System	This dataset includes a virtual physical process, a Python-based PLC simulation, a network simulation, and an HMI—all running on separate virtual machines and connected via Modbus/TCP.

WADI [133]	Cyber-Physical System	This dataset is a miniature water distribution testbed featuring 16 attacks designed to disrupt the water supply to consumer tanks.
Airsim [134]	Cyber-Physical System	This dataset is an open-source simulator that accommodates various vehicles, such as drones and cars.
TWOS [135]	User Identification and Authentication	This dataset includes both normal and malicious activities collected from various host-based heterogeneous sources such as mouse, keyboard, processes, and the file system.
Balabit[136]	User Identification and Authentication	This dataset includes the timing and positioning information of mouse pointers.
European Card [137]	Fraud detection	This dataset contains 284,807 transactions recorded over a few days, of which 492 are fraudulent.
IEEE-CIS [138]	Fraud detection	This real-world e-commerce transaction dataset includes 590,540 transactions, with 20,663 identified as fraud.
ISCXVPN 2016 [139]	Encrypted Traffic Analysis	This dataset includes 15 popular applications such as Facebook, YouTube, Vimeo, and Netflix, which encrypt traffic using various protocols.
Open HTTPS [140]	Encrypted Traffic Analysis	This dataset contains full HTTPS raw PCAP files collected from the top 779 accessed HTTPS websites.
UNB-CIC Tor[141]	Encrypted Traffic Analysis	This dataset stores various traffic types, including VOIP, chat, audio streaming, video streaming, mail, P2P, browsing, and File Transfer, from over 18 applications such as Spotify, Skype, and others.
DF-95 [142]	Encrypted Traffic Analysis	This dataset includes 95 websites, each containing 1,000 instances and up to 5000 cells per instance.
ASVSpoo f [143]	Malware detection	This dataset is used for evaluating audio spoofing detection and analyzing adversarial attacks on audio processing systems.
Fake or Real [144]	Android Malware	This dataset is designed to differentiate between fake and genuine audio samples, commonly used in anti-spoofing research.
ADD2023 [145].	Malware detection	This dataset is curated for the detection of audio deepfakes and features a wide variety of both real and synthetic audio samples.
In-The-Wild (ItW) [146].	Malware detection	This dataset is specifically created for real-world audio spoofing detection and includes a diverse range of challenging scenarios.
Ember Malware [147].	Android Malware	This dataset contains 1.1 million binary files, with 900,000 assigned for training and 200,000 for testing. The training part is evenly divided into 300,000 malicious, 300,000 harmless, and 300,000 unlabeled files. The testing set contains 100,000 malicious and 100,000 harmless samples.
STL-10 [148].	Image Classification	This dataset is designed for both unsupervised and semi-supervised learning, including 10 classes of labeled and unlabeled images

5.3.2 Benchmark Architectures

In this section we analyse different standard architectures used in adversarial attacks on deep learning and machine learning research, especially in the cybersecurity contexts within hybrid models, to evaluate their robustness against adversarial attacks. **Table 9** provides a summary of the most frequently referenced architectures in the literature.

Table 9: Common benchmark architecture

Ref	Description
ResNet-50 [149]	This 50-layer deep residual network is commonly used for image classification. Its skip connections help prevent vanishing gradients, but it remains vulnerable to adversarial attacks such as FGSM and PGD.
VGG-16 [150].	This 16-layer convolutional neural network (CNN) is appreciated for its simplicity and depth. However, its large number of parameters makes it more susceptible to adversarial perturbations.
GoogleNet (InceptionV1) [151].	This 22-layer network incorporates inception modules to lower computational costs without sacrificing accuracy. It has also been thoroughly tested for adversarial robustness.
MobileNet- V2 [152].	This lightweight CNN is designed for mobile and embedded vision applications. Its efficiency makes it vulnerable to resource-constrained adversarial attacks.
DenseNet121 [153].	This CNN has 121 heavily interconnected layers, where each layer is feed-forward linked to all others. While this improves feature reuse, it can also cause vulnerabilities in difficult situations.
BiLSTM (Bidirectional Long Short-Term Memory) [154].	This network merges memories from both directions and uses max-pooling for text classification, yet it remains vulnerable to adversarial attacks in NLP, like word-level perturbations.
BERT [155].	This is a pre-trained transformer-based language model, a crucial metric for evaluating NLP task performance. Despite its high effectiveness, BERT can still be vulnerable to adversarial attacks such as TextFooler and BERT-Attack.
PuVAE [156].	This model is a probabilistic variational autoencoder created to enhance robustness against adversarial attacks in generative tasks.

5.3.3 Evaluation Metrics

This section provides the evaluation metrics used to assess the defence methods used to mitigate adversarial attacks on ML and DL models in different applications. The main aspect of the evaluation of the defence strategies is how the attack can lower the system performance. Several research studies have assessed the impact of their methods on the model system across different algorithms and scenarios to demonstrate the effectiveness of the attacks [157] [158] [159] [160], [161] [162]. The selection of the parameter metrics is highly dependent on the types of attacks. Firstly, the metrics of FPR, TPR, accuracy, and precision are used to evaluate the poisoning attack on the DNN model [163] . The recall metric is used to assess the security of the deep learning model that is used to classify the network against adversarial perturbation attacks [164]. Secondly, the AUC and FPR metrics are recommended to be used in order to assess the pattern classification of the model [165][166]. Finally, **Table 10** provides a summary of the key performance metrics used to evaluate AI models.

Table 10: The most commonly used metrics for evaluating AI techniques and models.

Evaluation metric	Description	Equation
Confusion matrix	A 2×2 matrix listing False Positives (FP), False Negatives (FN), True Positives (TP) and True Negatives (TN).	none
False Positive Rate (FPR) or False Acceptance Rate (FAR)	The ratio of the elements wrongly determined as positives among the actual negatives.	$\frac{FP}{FP + TN}$
Recall, Sensitivity or True Positive Rate (TPR)	The ratio of the correctly identified positives.	$\frac{TP}{FN + TP}$
False Negative Rate (FNR) or False Rejection Rate (FRR)	The ratio of elements wrongly determined as negatives among the actual positives.	$\frac{FN}{FN + TP}$
Specificity or True Negative Rate (TNR)	The ratio of correctly identified negatives	$\frac{TN}{FP + TN}$
Precision	The proportion of actual positives among the elements predicted as positives.	$\frac{TP}{TP + FP}$
Accuracy	The proportion of correctly predicted items among the total number of items.	$\frac{TP + TN}{TP + TN + FP + FN}$
Accuracy Drop or Error Rate	This is an essential metric for assessing an AI model's robustness against adversarial and backdoor attacks	Error Rate $= \frac{N_{Incorrect\ Predictions}}{N_{Predictions}} \times 100\%$
Macro-Averaged Accuracy	This metric is a standard evaluation measure that calculates the average accuracy across all classes in a dataset.	Averaged_{Macro} (MAA) $= \frac{1}{C} \sum_{i=1}^C accuracy\ for\ class\ i$
F1-Score	The harmonic mean of precision and recall.	$\frac{2 \times precision \times recall}{precision + recall}$
Area Under Curve (AUC)	The area enclosed by the graph of TPR and FPR (ROC Curve) at different thresholds ranging from 0 to 1.	$\int ROC$
Attack Severity (AS)	The severity of the attack is inversely proportional to the detected attack samples.	$1 - \frac{Recall_{After\ attack}}{Recall_{Before\ attack}}$
Adversarial Attack Success Rate (ASR)	The proportion of misclassified adversarial examples among the total adversarial samples.	$FPR(X_{adversarial})$
Original Detection Rate (ODR)	The classifier's detection ability regarding legitimate samples.	$TNR(X_{original})$
Evasion Increase Rate (EIR)	The efficiency rate of the adversarial example generator.	$1 - \frac{ADR}{ODR}$
Average confidence($ACK(X)$)	X's average prediction confidence of the K class.	$\frac{1}{N_x} \sum_{i=1}^{N_x} p(k x_i)$
Average Confidence of Adversarial Class ($ACACK$)	Average prediction confidence of the misclassified samples of the K class	$AC_k(X_{Misclassified})$

AverageConfidenceofTrueClass ($ACTCk$)	Average prediction confidence of the correctly classified samples of the K class.	$AC_k(X_{CorrectClassified})$
TotalTimeCost (TTC)	Training and testing time consumption	$Time_{end}^{train} - Time_{start}^{Train} + Time_{end}^{Test} - Time_{start}^{Test}$
Effective Generation Rate (EGR), Evasion Rate (ER), Misclassification Rate (MR)	FPR of the adversarial examples.	$Avg_{max}(C(\hat{x})) \neq C_x, \quad x \approx \hat{x}$
Success Rate	The attack success rate (ASR) reflects the percentage of adversarial examples that successfully achieve their goal within the given attack budget	$ASR = \left(\frac{N_{Attempts}}{N_{Attacks}} \right) \times 100\%$
Average Success Rate (AvSR)	This is crucial metric to evaluate the effectiveness of continual learning techniques. It calculates the proportion of successful outcomes over various experiments or trials.	$AvSR = \frac{1}{N} \sum_{i=1}^n ASR_i$
Target Fooling Rate (TFR)	This specifies the event's occurrence rate, where a higher rate suggests that the classifier is more easily misled	$Avg_{max}(C(\hat{x})) = t \neq C_x, \quad x \approx \hat{x}$
Misclassification Rate	It also evaluates the proportion of instances that the model misclassifies, offering insights into its accuracy and reliability	$\frac{\text{misclassification Rate}}{FB + FN} = \frac{FB + FN}{TP + TN + FP + FN}$

6 Discussion

The findings of this survey for identifying the most critical adversarial attacks and their defended methods on the models of ML and DL have been presented in this section. The main challenge to integrating defence methods into AI models to mitigate adversarial attacks has been discussed accordingly. Future research directions in this important line of research have been discussed, while the practical implications for real-world implementation have been presented.

6.1 Key Challenges

Performance and scalability: This key challenge is considered the main matter that faces the defence method to mitigate adversarial attacks due to the success rates being decreased by 60%. This fact can significantly impact the false impression that model detection can burden system scalability in real-world implementation.

Defence Strategy Imbalances: This challenge highlights recent research that reveals notable differences, with 68% of defences focusing on detection and only 23% prioritizing training

improvement. Vulnerabilities arise because attackers can exploit flaws in the model creation process when it is still in its early stages of development.

Evaluation and Standardisation Gaps: This challenge shows that the current study found 47% of defence research techniques to be affected by this obstacle, highlighting issues with standardisation due to the difference in assessment frameworks. This uncertainty makes it harder to choose the best solutions and harder to compare how well different studies work for defence.

Integration Challenges: This challenge faced the AI model when it comes to combining different methods because each has its own operational and technical issues, such as zero-trust systems, associated learning, and quantum-resistant techniques. Privacy-preserving distributed defences often struggle to balance protecting privacy with maintaining strong security.

Implementation of defence strategy in Industry: The gap between academic research and industrial implementation of AI models is quite challenging. This is due to the lack of a standard checklist, which makes it more difficult to test defence methods.

6.2 Future Research Opportunities

This section outlines the future research directions of artificial intelligence in cybersecurity; multiple chances for future research have been highlighted in terms of technically complex and strategically significant aspects.

6.2.1 Robustness and Resilience of AI Models

The improvement of machine learning and deep learning models is a crucial priority due to the growing of refined adversarial attacks. A notable research direction should focus on developing a secure defence strategy that can effectively counter attacks like poisoning, white-box, and black-box.

6.2.2 Privacy-Preserving and Secure Learning

The combination of adversarial artificial intelligence (AI) in cybersecurity offers great privacy-preserving features. However, the growing trends of federated learning utilisation, differential privacy, and secure multi-party computation in the field of cybersecurity systems are promising to

improve the AI model against adversarial attacks. This line of research can mitigate data leakage during model training.

6.2.3 AI-Driven Decision Support Systems for Cybersecurity

The development of adversarial attacks led to misclassification in the AI model, resulting in reduced accuracy in detecting these attacks. On the other hand, the development of Decision Support Systems (DSS) on an AI model offers promising improvement for their system. Notable research directions should focus on integrating DSS with a new state-of-the-art defence of AI models to ensure resistance against refined threats, such as evasion, poisoning, and model reversal.

6.2.4 Integration of Blockchain for Secure AI Systems

The characteristics of blockchain, such as decentralisation, integrity, and transparency, offer resilience to federated learning in AI models. One promising future direction is to handle the potential of blockchain in defending federated learning against adversarial attacks.

6.2.5 5G and Edge Computing for Real-Time AI Defence

The development of 5G and edge computing ensures the security of the data generated by sensors from possible attacks. The integration of an AI model with an intrusion detection system is used to prevent these attacks. However, researchers should focus on improving the AI model by implementing a new defence strategy to ensure its resilience against adversarial attacks. Also, it is important to maintain the computational complexity of the model, especially in a limited-resource environment.

6.2.6 Global Standardisation of AI Security Frameworks

When it comes to implementing an effective and secure AI model, a standard framework is needed to assess the model's effectiveness. However, future research is needed to establish widely accepted procedures and standards for assessing AI models in terms of adversarial attacks and their defence strategies.

6.3 Practical Implications

The development of artificial intelligence models, such as machine learning (ML) and deep learning (DL) models, remains vulnerable to adversarial attacks. Researchers are required to use adversarial training defence strategies in addition to the traditional detection system due to balance between security and performance.

Also, the developed taxonomy of the current study is a good framework for cybersecurity experts to evaluate the security of AI models against various types of attacks. The developer can also evaluate their system, such as intrusion detection, malware classification, or biometric authentication, if required to adopt a defence strategy.

In real-world cybersecurity applications, such as encrypted traffic analysis, autonomous systems, and the Internet of Things, it is important to take into consideration the performance of the detection system in terms of scalability, latency, and resource limits. Certain of the defence strategies require high computational complexity and are not suitable for some cybersecurity applications. However, the current study provided a tremendous framework that allows the developer to assess their proposed solution.

Standardised assessment frameworks play a crucial role in developing cybersecurity applications across the industry. Researchers, policymakers, and regulators should collaborate to establish standards for model validation, adversarial testing, and the ethical deployment of models. Such cooperation will improve the trust in AI systems, enabling their safe integration into organisational and national cybersecurity strategies.

7. Conclusion

The limitation of the current study is the database used in our inquiry search. The rapid increase in research on adversarial attacks in DL and ML models can affect the timeline of the study. Adversarial attacks are a significant threat to cybersecurity applications, since they render high-performance ML/DL models vulnerable when targeting them. This study explores key gaps by reviewing 1,000 publications from 2017 to 2025 that show a 40% increase in defence-related research since 2020. This study has a significant impact on the development of AI models in different applications. The following concluding statements are constructed:

- A new three-level taxonomy that categorises adversarial attacks based on knowledge, timing, goals, and types, which provides a clearer framework than the earlier models; creative defence classification highlights strategies such as federated learning, zero-trust architectures, and quantum-resistant approaches.
- The integrated adversarial attacks map introduces 12 new attack combinations, and data indicates that combined defences improve protection by 15-25% when compared to individual measures.
- The conducted research noted that 68% of defence methods are mainly focused on detection, while adversarial attacks decrease by 60% when moving from laboratory to real-world settings.
- This study discussed the challenges that can be faced in AI models in the aspect of adversarial attacks in different applications. Future research opportunities and prospects are available for researchers and organisations in many aspects.

Statements and Declarations

Funding: The authors declare that no funds, grants, or other support were received during the preparation of this manuscript.

Competing Interests: The authors have no relevant financial or non-financial interests to disclose.

Author Contributions: The authors contributed to the study conception, numerical modelling, and data analysis. All authors read and approved the final manuscript.

Data Availability: The datasets generated during the current study are available from the corresponding author on reasonable request.

Clinical Trial Registration: Not applicable.

Consent to Publish: Not applicable.

Consent to Participate: Not applicable

Ethics Declaration: Not applicable.

References

- [1] A. H. Salem, S. M. Azzam, O. E. Emam, and A. A. Abohany, "Advancing cybersecurity: a comprehensive review of AI-driven detection techniques," *J Big Data*, vol. 11, no. 1, Dec. 2024, doi: 10.1186/s40537-024-00957-y.
- [2] I. Azhar, "How Artificial Intelligence Is Changing Cyber Security Landscape and Preventing Cyber Attacks: A systematic review," *Ishaq Azhar Mohammed," HOW ARTIFICIAL INTELLIGENCE IS CHANGING CYBER SECURITY LANDSCAPE AND PREVENTING CYBER ATTACKS: A SYSTEMATIC REVIEW"*, *International Journal of Creative Research Thoughts (IJCRT)*, ISSN, pp. 2320–2882, 2016.
- [3] M. S. Akhtar and T. Feng, "Malware analysis and detection using machine learning algorithms," *Symmetry (Basel)*, vol. 14, no. 11, p. 2304, 2022.
- [4] G. Lingam, R. R. Rout, and D. V. L. N. Somayajulu, "Adaptive deep Q-learning model for detecting social bots and influential users in online social networks," *Applied Intelligence*, vol. 49, no. 11, pp. 3947–3964, 2019.
- [5] A. Graves, "Generating sequences with recurrent neural networks," *arXiv preprint arXiv:1308.0850*, 2013.
- [6] I. Goodfellow *et al.*, "Generative adversarial networks," *Commun ACM*, vol. 63, no. 11, pp. 139–144, 2020.
- [7] P. Vincent, H. Larochelle, I. Lajoie, Y. Bengio, P.-A. Manzagol, and L. Bottou, "Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion.," *Journal of machine learning research*, vol. 11, no. 12, 2010.
- [8] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning representations by back-propagating errors," *Nature*, vol. 323, no. 6088, pp. 533–536, 1986.
- [9] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.
- [10] S. Okdem and S. Okdem, "Artificial Intelligence in Cybersecurity: A Review and a Case Study," *Applied Sciences (Switzerland)*, vol. 14, no. 22, Nov. 2024, doi: 10.3390/app142210487.
- [11] G. G. Shaye, M. H. M. Zabil, M. A. Habeeb, Y. L. Khaleel, and A. S. Albahri, "Strategies for protection against adversarial attacks in AI models: An in-depth review," Jan. 01, 2025, *Walter de Gruyter GmbH*. doi: 10.1515/jisys-2024-0277.

- [12] M. Goldblum *et al.*, “Dataset security for machine learning: Data poisoning, backdoor attacks, and defenses,” *IEEE Trans Pattern Anal Mach Intell*, vol. 45, no. 2, pp. 1563–1580, 2022.
- [13] N. AI, “Artificial intelligence risk management framework: Generative artificial intelligence profile,” *NIST Trustworthy and Responsible AI Gaithersburg, MD, USA*, 2024.
- [14] X. Yuan, P. He, Q. Zhu, and X. Li, “Adversarial examples: Attacks and defenses for deep learning,” *IEEE Trans Neural Netw Learn Syst*, vol. 30, no. 9, pp. 2805–2824, 2019.
- [15] N. Akhtar and A. Mian, “Threat of adversarial attacks on deep learning in computer vision: A survey,” *Ieee Access*, vol. 6, pp. 14410–14430, 2018.
- [16] F. O. Olowononi, D. B. Rawat, and C. Liu, “Resilient machine learning for networked cyber physical systems: A survey for machine learning security to securing machine learning for CPS,” *IEEE Communications Surveys & Tutorials*, vol. 23, no. 1, pp. 524–552, 2020.
- [17] I. Rosenberg, A. Shabtai, Y. Elovici, and L. Rokach, “Adversarial machine learning attacks and defense methods in the cyber security domain,” *ACM Computing Surveys (CSUR)*, vol. 54, no. 5, pp. 1–36, 2021.
- [18] S.-M. Moosavi-Dezfooli, A. Fawzi, and P. Frossard, “Deepfool: a simple and accurate method to fool deep neural networks,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2574–2582.
- [19] I. J. Goodfellow, J. Shlens, and C. Szegedy, “Explaining and harnessing adversarial examples,” *arXiv preprint arXiv:1412.6572*, 2014.
- [20] N. Carlini and D. Wagner, “Towards evaluating the robustness of neural networks,” in *2017 IEEE Symposium on Security and Privacy (SP)*, Ieee, 2017, pp. 39–57.
- [21] A. Kurakin, I. Goodfellow, and S. Bengio, “Adversarial machine learning at scale,” *arXiv preprint arXiv:1611.01236*, 2016.
- [22] Ö. Aslan, S. S. Aktuğ, M. Ozkan-Okay, A. A. Yilmaz, and E. Akin, “A comprehensive review of cyber security vulnerabilities, threats, attacks, and solutions,” *Electronics (Basel)*, vol. 12, no. 6, p. 1333, 2023.
- [23] G. Bathla *et al.*, “Autonomous vehicles and intelligent automation: Applications, challenges, and opportunities,” *Mobile Information Systems*, vol. 2022, no. 1, p. 7632892, 2022.

- [24] M. Girdhar, J. Hong, and J. Moore, "Cybersecurity of autonomous vehicles: A systematic literature review of adversarial attacks and defense models," *IEEE Open Journal of Vehicular Technology*, vol. 4, pp. 417–437, 2023.
- [25] N. AI, "Artificial intelligence risk management framework: Generative artificial intelligence profile," *NIST Trustworthy and Responsible AI Gaithersburg, MD, USA*, 2024.
- [26] S. H. Silva and P. Najafirad, "Opportunities and challenges in deep learning adversarial robustness: A survey," *arXiv preprint arXiv:2007.00753*, 2020.
- [27] M. Girdhar, J. Hong, and J. Moore, "Cybersecurity of autonomous vehicles: A systematic literature review of adversarial attacks and defense models," *IEEE Open Journal of Vehicular Technology*, vol. 4, pp. 417–437, 2023.
- [28] T. Abbas, "Adversarial AI in Cybersecurity: How Machine Learning Can Both Attack and Defend Digital Systems," *Authorea Preprints*, 2025.
- [29] Y. Wang *et al.*, "Adversarial Attacks and Defenses in Machine Learning-Powered Networks: A Contemporary Survey," Mar. 2023, [Online]. Available: <http://arxiv.org/abs/2303.06302>
- [30] S. Okdem and S. Okdem, "Artificial Intelligence in Cybersecurity: A Review and a Case Study," *Applied Sciences (Switzerland)*, vol. 14, no. 22, Nov. 2024, doi: 10.3390/app142210487.
- [31] H. Mustafa, A. U. Soykat, R. Rahman, and M. B. Biplob, "A Comprehensive Review of Large Language Models and AI in Cybersecurity: Applications in Threat Detection, Defense, and Software Security," 2025.
- [32] W. Jiang, Z. He, J. Zhan, W. Pan, and D. Adhikari, "Research progress and challenges on application-driven adversarial examples: A survey," *ACM Transactions on Cyber-Physical Systems (TCPS)*, vol. 5, no. 4, pp. 1–25, 2021.
- [33] J. Vadillo, R. Santana, and J. A. Lozano, "Adversarial attacks in explainable machine learning: A survey of threats against models and humans," *Wiley Interdiscip Rev Data Min Knowl Discov*, vol. 15, no. 1, p. e1567, 2025.
- [34] A. Kuzior and M. Sira, "A bibliometric analysis of blockchain technology research using VOSviewer," *Sustainability*, vol. 14, no. 13, p. 8206, 2022.
- [35] N. Donthu, G. Kumar Badhotiya, S. Kumar, G. Soni, and N. Pandey, "A retrospective overview of Journal of Enterprise Information Management using bibliometric analysis," *Journal of Enterprise Information Management*, vol. 35, no. 2, pp. 504–529, 2022.

- [36] H. M. Hussien, S. M. Yasin, N. I. Udzir, M. I. H. Ninggal, and S. Salman, "Blockchain technology in the healthcare industry: Trends and opportunities," *J Ind Inf Integr*, vol. 22, p. 100217, 2021.
- [37] H. J. Akeiber, "A comprehensive study of Cybercrime and Digital Forensics through Machine Learning and AI," *Al-Rafidain Journal of Engineering Sciences*, pp. 369–395, 2025.
- [38] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," *arXiv preprint arXiv:1706.06083*, 2017.
- [39] O. Ibitoye, O. Shafiq, and A. Matrawy, "Analyzing adversarial attacks against deep learning for intrusion detection in IoT networks," in *2019 IEEE global communications conference (GLOBECOM)*, IEEE, 2019, pp. 1–6.
- [40] H. Qiu, T. Dong, T. Zhang, J. Lu, G. Memmi, and M. Qiu, "Adversarial attacks against network intrusion detection in IoT systems," *IEEE Internet Things J*, vol. 8, no. 13, pp. 10327–10335, 2020.
- [41] F. O. Olowononi, D. B. Rawat, and C. Liu, "Resilient machine learning for networked cyber physical systems: A survey for machine learning security to securing machine learning for CPS," *IEEE Communications Surveys & Tutorials*, vol. 23, no. 1, pp. 524–552, 2020.
- [42] A. Rawal, J. McCoy, D. B. Rawat, B. M. Sadler, and R. S. Amant, "Recent advances in trustworthy explainable artificial intelligence: Status, challenges, and perspectives," *IEEE Transactions on Artificial Intelligence*, vol. 3, no. 6, pp. 852–866, 2021.
- [43] M. Pawlicki, M. Choraś, and R. Kozik, "Defending network intrusion detection systems against adversarial evasion attacks," *Future Generation Computer Systems*, vol. 110, pp. 148–154, 2020.
- [44] Y. Deng, T. Zhang, G. Lou, X. Zheng, J. Jin, and Q.-L. Han, "Deep learning-based autonomous driving systems: A survey of attacks and defenses," *IEEE Trans Industr Inform*, vol. 17, no. 12, pp. 7897–7912, 2021.
- [45] E. N. Crothers, N. Japkowicz, and H. L. Viktor, "Machine-generated text: A comprehensive survey of threat models and detection methods," *IEEE Access*, vol. 11, pp. 70977–71002, 2023.
- [46] A. Kuppa and N.-A. Le-Khac, "Adversarial XAI methods in cybersecurity," *IEEE transactions on information forensics and security*, vol. 16, pp. 4924–4938, 2021.

- [47] F. A. Yerlikaya and Ş. Bahtiyar, “Data poisoning attacks against machine learning algorithms,” *Expert Syst Appl*, vol. 208, p. 118101, 2022.
- [48] A. Mohammed, “Artificial Intelligence-Powered Cyber Attacks: Adversarial Machine Learning,” *Authorea Preprints*, 2025.
- [49] A. Alotaibi and M. A. Rassam, “Adversarial machine learning attacks against intrusion detection systems: A survey on strategies and defense,” *Future Internet*, vol. 15, no. 2, p. 62, 2023.
- [50] H. Feng, S. Li, H. Shi, and Z. Ye, “A Comparative Analysis of White Box and Gray Box Adversarial Attacks to Natural Language Processing Systems,” in *2024 2nd International Conference on Image, Algorithms and Artificial Intelligence (ICIAAI 2024)*, Atlantis Press, 2024, pp. 640–646.
- [51] A. Mohammed, “Artificial Intelligence-Powered Cyber Attacks: Adversarial Machine Learning,” *Authorea Preprints*, 2025.
- [52] O. M. Ijiga, I. P. Idoko, G. I. Ebiega, F. I. Olajide, T. I. Olatunde, and C. Ukaegbu, “Harnessing adversarial machine learning for advanced threat detection: AI-driven strategies in cybersecurity risk assessment and fraud prevention,” *J. Sci. Technol*, vol. 11, pp. 1–24, 2024.
- [53] J. C. Costa, T. Roxo, H. Proença, and P. R. M. Inacio, “How deep learning sees the world: A survey on adversarial attacks & defenses,” *IEEE Access*, vol. 12, pp. 61113–61136, 2024.
- [54] U. A. Ali, K. Dogra, and S. Sharma, “White-box adversarial exploitation of NIDS: Insights from FGSM, PGD, and C&W,” in *2025 2nd International Conference on Computational Intelligence, Communication Technology and Networking (CICTN)*, IEEE, 2025, pp. 668–673.
- [55] S. Agnihotri, S. Jung, and M. Keuper, “CosPGD: an efficient white-box adversarial attack for pixel-wise prediction tasks,” *arXiv preprint arXiv:2302.02213*, 2023.
- [56] Y. Zhang and P. Liang, “Defending against whitebox adversarial attacks via randomized discretization,” in *The 22nd International Conference on Artificial Intelligence and Statistics*, PMLR, 2019, pp. 684–693.
- [57] X. Zhong and C. Liu, “Sparse-PGD: A Unified Framework for Sparse Adversarial Perturbations Generation,” *arXiv preprint arXiv:2405.05075*, 2024.

- [58] L. Ye and S. M. Hamidi, “Thundernna: a white box adversarial attack,” *arXiv preprint arXiv:2111.12305*, 2021.
- [59] C. T. Lei, Z. Guo, H. C. Lee, M. Q. Duong, and C. P. Lau, “Towards more transferable adversarial attack in black-box manner,” *arXiv preprint arXiv:2505.18097*, 2025.
- [60] S. Agnihotri, S. Jung, and M. Keuper, “CosPGD: an efficient white-box adversarial attack for pixel-wise prediction tasks,” *arXiv preprint arXiv:2302.02213*, 2023.
- [61] S. Agnihotri, S. Jung, and M. Keuper, “CosPGD: an efficient white-box adversarial attack for pixel-wise prediction tasks,” *arXiv preprint arXiv:2302.02213*, 2023.
- [62] Y. Li, M. Cheng, C.-J. Hsieh, and T. C. M. Lee, “A review of adversarial attack and defense for classification methods,” *Am Stat*, vol. 76, no. 4, pp. 329–345, 2022.
- [63] W. Fan *et al.*, “Adversarial attacks for black-box recommender systems via copying transferable cross-domain user profiles,” *IEEE Trans Knowl Data Eng*, vol. 35, no. 12, pp. 12415–12429, 2023.
- [64] K. Tan, J. Wang, and Y. Kantaros, “Targeted adversarial attacks against neural network trajectory predictors,” in *Learning for Dynamics and Control Conference*, PMLR, 2023, pp. 431–444.
- [65] H. Ren and T. Huang, “Adversarial example attacks in the physical world,” in *International Conference on Machine Learning for Cyber Security*, Springer, 2020, pp. 572–582.
- [66] M. Standen, J. Kim, and C. Szabo, “Adversarial machine learning attacks and defences in multi-agent reinforcement learning,” *ACM Comput Surv*, vol. 57, no. 5, pp. 1–35, 2025.
- [67] L. Schwinn, R. Raab, A. Nguyen, D. Zanca, and B. Eskofier, “Exploring misclassifications of robust neural networks to enhance adversarial attacks,” *Applied intelligence*, vol. 53, no. 17, pp. 19843–19859, 2023.
- [68] E. R. Balda, A. Behboodi, and R. Mathar, “Perturbation analysis of learning algorithms: Generation of adversarial examples from classification to regression,” *IEEE Transactions on Signal Processing*, vol. 67, no. 23, pp. 6078–6091, 2019.
- [69] Z. Wang, F. Nie, H. Wang, H. Huang, and F. Wang, “Toward robust discriminative projections learning against adversarial patch attacks,” *IEEE Trans Neural Netw Learn Syst*, 2023.
- [70] M. Abroshan and S.-M. Moosavi-Dezfooli, “SuperDeepFool: a new fast and accurate minimal adversarial attack,” *Adv Neural Inf Process Syst*, vol. 37, pp. 98537–98562, 2024.

- [71] S.-M. Moosavi-Dezfooli, A. Fawzi, and P. Frossard, “Deepfool: a simple and accurate method to fool deep neural networks,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2574–2582.
- [72] P. Cheng *et al.*, “UniAP: Protecting speech privacy with non-targeted universal adversarial perturbations,” *IEEE Trans Dependable Secure Comput*, vol. 21, no. 1, pp. 31–46, 2023.
- [73] Y. Wang *et al.*, “Adversarial attacks and defenses in machine learning-empowered communication systems and networks: A contemporary survey,” *IEEE Communications Surveys & Tutorials*, vol. 25, no. 4, pp. 2245–2298, 2023.
- [74] W. Villegas-Ch, A. Jaramillo-Alcázar, and S. Luján-Mora, “Evaluating the robustness of deep learning models against adversarial attacks: An analysis with fgsm, pgd and cw,” *Big Data and Cognitive Computing*, vol. 8, no. 1, p. 8, 2024.
- [75] R. R. Kethireddy, “Adversarial Machine Learning: Attacks and Defense Mechanisms with Respect to AI Security”.
- [76] E. Charles and P. Krishnan, “Adversarial Attacks in Deep Learning: Analyzing Vulnerabilities and Designing Robust Defense Mechanisms,” 2025.
- [77] O. S. Ajibola, A. Adetoye, O. O. Awoyomi, and S. C. Mweni, “Countermeasures for Adversarial AI Attacks in Cybersecurity: Open Challenges and Future Research Directions”.
- [78] O. S. Ajibola, A. Adetoye, O. O. Awoyomi, and S. C. Mweni, “Countermeasures for Adversarial AI Attacks in Cybersecurity: Open Challenges and Future Research Directions”.
- [79] T. Kim, H. J. Lee, and Y. M. Ro, “Map: Multispectral adversarial patch to attack person detection,” in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2022, pp. 4853–4857.
- [80] S. Chen, Z. He, C. Sun, J. Yang, and X. Huang, “Universal adversarial attack on attention and the resulting dataset damagenet,” *IEEE Trans Pattern Anal Mach Intell*, vol. 44, no. 4, pp. 2188–2197, 2020.
- [81] C. Luo, Q. Lin, W. Xie, B. Wu, J. Xie, and L. Shen, “Frequency-driven imperceptible adversarial attack on semantic similarity,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 15315–15324.

- [82] J. Li, T. Xie, L. Chen, F. Xie, X. He, and Z. Zheng, “Adversarial attack on large scale graph,” *IEEE Trans Knowl Data Eng*, vol. 35, no. 1, pp. 82–95, 2021.
- [83] J. Chen, X. Lin, Z. Shi, and Y. Liu, “Link prediction adversarial attack via iterative gradient attack,” *IEEE Trans Comput Soc Syst*, vol. 7, no. 4, pp. 1081–1094, 2020.
- [84] H. Chang *et al.*, “Adversarial attack framework on graph embedding models with limited knowledge,” *IEEE Trans Knowl Data Eng*, vol. 35, no. 5, pp. 4499–4513, 2022.
- [85] S. Chen, Z. Chen, and D. Wang, “Adaptive adversarial training for meta reinforcement learning,” in *2021 International Joint Conference on Neural Networks (IJCNN)*, IEEE, 2021, pp. 1–8.
- [86] J. Wang, M. Luo, F. Suya, J. Li, Z. Yang, and Q. Zheng, “Scalable attack on graph data by injecting vicious nodes,” *Data Min Knowl Discov*, vol. 34, no. 5, pp. 1363–1389, 2020.
- [87] I. J. Goodfellow, J. Shlens, and C. Szegedy, “Explaining and harnessing adversarial examples,” *arXiv preprint arXiv:1412.6572*, 2014.
- [88] B. Biggio *et al.*, “Evasion attacks against machine learning at test time,” in *Joint European conference on machine learning and knowledge discovery in databases*, Springer, 2013, pp. 387–402.
- [89] T. S. Sethi and M. Kantardzic, “Data driven exploratory attacks on black box classifiers in adversarial domains,” *Neurocomputing*, vol. 289, pp. 129–143, 2018.
- [90] Y.-Y. Ding, H.-J. Lin, L.-J. Liu, Z.-H. Ling, and Y. Hu, “Robustness of speech spoofing detectors against adversarial post-processing of voice conversion,” *IEEE/ACM Trans Audio Speech Lang Process*, vol. 29, pp. 3415–3426, 2021.
- [91] A. Liu *et al.*, “Perceptual-sensitive gan for generating adversarial patches,” in *Proceedings of the AAAI conference on artificial intelligence*, 2019, pp. 1028–1035.
- [92] K. Han, Y. Li, and B. Xia, “A cascade model-aware generative adversarial example detection method,” *Tsinghua Sci Technol*, vol. 26, no. 6, pp. 800–812, 2021.
- [93] S. Freitas, S.-T. Chen, Z. J. Wang, and D. H. Chau, “Unmask: Adversarial detection and defense through robust feature alignment,” in *2020 IEEE International Conference on Big Data (Big Data)*, IEEE, 2020, pp. 1081–1088.
- [94] X. Zhang and Y. Wang, “ADNet: A neural network model for adversarial example detection based on steganalysis and attention mechanism,” in *2022 3rd International Conference on*

- Computer Vision, Image and Deep Learning & International Conference on Computer Engineering and Applications (CVIDL & ICCEA)*, IEEE, 2022, pp. 55–60.
- [95] G. Chen, P. Peng, L. Ma, J. Li, L. Du, and Y. Tian, “Amplitude-phase recombination: Rethinking robustness of convolutional neural networks in frequency domain,” in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 458–467.
 - [96] Y. Li *et al.*, “Towards robustness of deep neural networks via regularization,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 7496–7505.
 - [97] N. Papernot, P. McDaniel, X. Wu, S. Jha, and A. Swami, “Distillation as a defense to adversarial perturbations against deep neural networks,” in *2016 IEEE symposium on security and privacy (SP)*, IEEE, 2016, pp. 582–597.
 - [98] F. O. Catak, M. Kuzlu, E. Catak, U. Cali, and O. Guler, “Defensive distillation-based adversarial attack mitigation method for channel estimation using deep learning models in next-generation wireless networks,” *IEEE Access*, vol. 10, pp. 98191–98203, 2022.
 - [99] H. C. Moon, S. Joty, and X. Chi, “GradMask: Gradient-guided token masking for textual adversarial example detection,” in *Proceedings of the 28th ACM SIGKDD conference on knowledge discovery and data mining*, 2022, pp. 3603–3613.
 - [100] H. Lee, H. Bae, and S. Yoon, “Gradient masking of label smoothing in adversarial robustness,” *IEEE Access*, vol. 9, pp. 6453–6464, 2020.
 - [101] X. Yan, Y. Li, T. Dai, Y. Bai, and S.-T. Xia, “D2Defend: Dual-domain based defense against adversarial examples,” in *2021 International Joint Conference on Neural Networks (IJCNN)*, IEEE, 2021, pp. 1–8.
 - [102] B. Zi, S. Zhao, X. Ma, and Y.-G. Jiang, “Revisiting adversarial robustness distillation: Robust soft labels make student better,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 16443–16452.
 - [103] Z. Yan, Y. Guo, and C. Zhang, “Adversarial margin maximization networks,” *IEEE Trans Pattern Anal Mach Intell*, vol. 43, no. 4, pp. 1129–1139, 2019.
 - [104] C. Xiang, S. Mahloujifar, and P. Mittal, “{PatchCleanser}: Certifiably robust defense against adversarial patches for any image classifier,” in *31st USENIX security symposium (USENIX Security 22)*, 2022, pp. 2065–2082.

- [105] P. Chiang, R. Ni, A. Abdelkader, C. Zhu, C. Studer, and T. Goldstein, “Certified defenses for adversarial patches,” *arXiv preprint arXiv:2003.06693*, 2020.
- [106] T. Bai, J. Luo, J. Zhao, B. Wen, and Q. Wang, “Recent advances in adversarial training for adversarial robustness,” *arXiv preprint arXiv:2102.01356*, 2021.
- [107] H. Zhang and J. Wang, “Defense against adversarial attacks using feature scattering-based adversarial training,” *Adv Neural Inf Process Syst*, vol. 32, 2019.
- [108] A. Shafahi *et al.*, “Adversarial training for free!,” *Adv Neural Inf Process Syst*, vol. 32, 2019.
- [109] W. Jin *et al.*, “Adversarial attacks and defenses on graphs,” *ACM SIGKDD Explorations Newsletter*, vol. 22, no. 2, pp. 19–34, 2021.
- [110] F. Feng, X. He, J. Tang, and T.-S. Chua, “Graph adversarial training: Dynamically regularizing based on graph structure,” *IEEE Trans Knowl Data Eng*, vol. 33, no. 6, pp. 2493–2504, 2019.
- [111] X. Liu *et al.*, “Mape: defending against transferable adversarial attacks using multi-source adversarial perturbations elimination,” *Complex & Intelligent Systems*, vol. 11, no. 2, p. 155, 2025.
- [112] H. Kuang, H. Liu, X. Lin, and R. Ji, “Defense against adversarial attacks using topology aligning adversarial training,” *IEEE Transactions on Information Forensics and Security*, vol. 19, pp. 3659–3673, 2024.
- [113] H. Yan, J. Li, L. Du, B. Fang, Y. Jia, and Z. Gu, “Adversarial Hierarchical-Aware Edge Attention Learning Method for Network Intrusion Detection,” *Applied Sciences*, vol. 15, no. 14, p. 7915, 2025.
- [114] G. Sriramanan, S. Addepalli, and A. Baburaj, “Guided adversarial attack for evaluating and enhancing adversarial defenses,” *Adv Neural Inf Process Syst*, vol. 33, pp. 20297–20308, 2020.
- [115] T. Strauss, M. Hanselmann, A. Junginger, and H. Ulmer, “Ensemble methods as a defense to adversarial perturbations against deep neural networks,” *arXiv preprint arXiv:1709.03423*, 2017.
- [116] Y.-D. Lin *et al.*, “ELAT: Ensemble learning with adversarial training in defending against evaded intrusions,” *Journal of Information Security and Applications*, vol. 71, p. 103348, 2022.

- [117] Z. Lu, H. Hu, S. Huo, and S. Li, "Ensemble learning methods of adversarial attacks and defenses in computer vision: Recent progress," in *2021 International Conference on Advanced Computing and Endogenous Security*, IEEE, 2022, pp. 1–10.
- [118] J. Chen, C. Yuan, J. Li, D. Tian, R. Ma, and X. Jia, "ELAMD: An ensemble learning framework for adversarial malware defense," *Journal of Information Security and Applications*, vol. 75, p. 103508, 2023.
- [119] E. Nowroozi, M. Mohammadi, E. Savaş, Y. Mekdad, and M. Conti, "Employing deep ensemble learning for improving the security of computer networks against adversarial attacks," *IEEE Transactions on Network and Service Management*, vol. 20, no. 2, pp. 2096–2105, 2023.
- [120] K. Allix, T. F. Bissyandé, J. Klein, and Y. Le Traon, "Androzoo: Collecting millions of android apps for the research community," in *Proceedings of the 13th international conference on mining software repositories*, 2016, pp. 468–471.
- [121] Y. M. P. Pa, S. Suzuki, K. Yoshioka, T. Matsumoto, T. Kasama, and C. Rossow, "IoTPOT: A novel honeypot for revealing current IoT threats," *Journal of Information Processing*, vol. 24, no. 3, pp. 522–533, 2016.
- [122] Y. Zhou and X. Jiang, "Dissecting android malware: Characterization and evolution," in *2012 IEEE symposium on security and privacy*, IEEE, 2012, pp. 95–109.
- [123] D. Arp, M. Spreitzenbarth, M. Hubner, H. Gascon, K. Rieck, and C. Siemens, "Drebin: Effective and explainable detection of android malware in your pocket.," in *Ndss*, 2014, pp. 23–26.
- [124] Plohmann D., "Dgarchive: Domain names that are algorithmically created by malware using DGAs." Accessed: Aug. 11, 2025. [Online]. Available: <https://dgarchive.caad.fkie.fraunhofer.de/>
- [125] M. Zago, M. G. Pérez, and G. M. Pérez, "UMUDGA: A dataset for profiling DGA-based botnet," *Comput Secur*, vol. 92, p. 101719, 2020.
- [126] N. Koroniotis, N. Moustafa, E. Sitnikova, and B. Turnbull, "Towards the development of realistic botnet dataset in the internet of things for network forensic analytics: Bot-iot dataset," *Future Generation Computer Systems*, vol. 100, pp. 779–796, 2019.

- [127] I. Sharafaldin, A. H. Lashkari, and A. A. Ghorbani, "Toward generating a new intrusion detection dataset and intrusion traffic characterization.," *ICISSp*, vol. 1, no. 2018, pp. 108–116, 2018.
- [128] I. Sharafaldin, A. H. Lashkari, S. Hakak, and A. A. Ghorbani, "Developing realistic distributed denial of service (DDoS) attack dataset and taxonomy," in *2019 international carnahhan conference on security technology (ICCST)*, IEEE, 2019, pp. 1–8.
- [129] L. A. H. , G. A. A. Sharafaldin I., "A realistic cyber defense dataset (CSE-CIC-IDS2018)." Accessed: Aug. 11, 2025. [Online]. Available: <https://www.unb.ca/cic/datasets/ids-2018.html>
- [130] A. Guerra-Manzanares, J. Medina-Galindo, H. Bahsi, and S. Nõmm, "MedBIoT: Generation of an IoT botnet dataset in a medium-sized IoT network.," in *ICISSP*, 2020, pp. 207–218.
- [131] M. Aghashahi, R. Sundararajan, M. Pourahmadi, and M. K. Banks, "Water distribution systems analysis symposium–battle of the attack detection algorithms (batadal)," in *World environmental and water resources congress 2017*, 2017, pp. 101–108.
- [132] T. H. Morris, Z. Thornton, and I. Turnipseed, "Industrial control system simulation and data logging for intrusion detection system research," *7th annual southeastern cyber security summit*, pp. 3–4, 2015.
- [133] C. M. Ahmed, V. R. Palleti, and A. P. Mathur, "WADI: a water distribution testbed for research in the design of secure cyber physical systems," in *Proceedings of the 3rd international workshop on cyber-physical systems for smart water networks*, 2017, pp. 25–28.
- [134] S. Shah, D. Dey, C. Lovett, and A. Kapoor, "Airsim: High-fidelity visual and physical simulation for autonomous vehicles," in *Field and service robotics: Results of the 11th international conference*, Springer, 2017, pp. 621–635.
- [135] A. Harilal, F. Toffalini, J. Castellanos, J. Guarnizo, I. Homoliak, and M. Ochoa, "Twos: A dataset of malicious insider threat behavior based on a gamified competition," in *Proceedings of the 2017 international workshop on managing insider security threats*, 2017, pp. 45–56.
- [136] K. L. , K. T. , W.-P. E. Fülöp A., "Mouse dynamics challenge." Accessed: Aug. 11, 2025. [Online]. Available: <https://github.com/balabit/Mouse-Dynamics-Challenge>

- [137] A. Dal Pozzolo, O. Caelen, R. A. Johnson, and G. Bontempi, “Calibrating probability with undersampling for unbalanced classification,” in *2015 IEEE symposium series on computational intelligence*, IEEE, 2015, pp. 159–166.
- [138] Vesta Corporation G., “IEEE-CIS fraud detection.” Accessed: Aug. 11, 2025. [Online]. Available: <https://www.kaggle.com/c/ieee-fraud-detection>
- [139] G. Draper-Gil, A. H. Lashkari, M. S. I. Mamun, and A. A. Ghorbani, “Characterization of encrypted and vpn traffic using time-related,” in *Proceedings of the 2nd international conference on information systems security and privacy (ICISSP)*, 2016, pp. 407–414.
- [140] T. C. , J. F. , I. C. Wazen S., “Open HTTPS Dataset.” Accessed: Aug. 11, 2025. [Online]. Available: <https://betternet.lhs.loria.fr/datasets/https/index.html>
- [141] A. H. Lashkari, G. D. Gil, M. S. I. Mamun, and A. A. Ghorbani, “Characterization of tor traffic using time based features,” in *International conference on information systems security and privacy*, SciTePress, 2017, pp. 253–262.
- [142] P. Sirinam, M. Imani, M. Juarez, and M. Wright, “Deep fingerprinting: Undermining website fingerprinting defenses with deep learning,” in *Proceedings of the 2018 ACM SIGSAC conference on computer and communications security*, 2018, pp. 1928–1943.
- [143] X. Wang *et al.*, “ASVspoof 2019: A large-scale public database of synthesized, converted and replayed speech,” *Comput Speech Lang*, vol. 64, p. 101114, 2020.
- [144] R. Reimao and V. Tzerpos, “For: A dataset for synthetic speech detection,” in *2019 International Conference on Speech Technology and Human-Computer Dialogue (SpeD)*, IEEE, 2019, pp. 1–10.
- [145] J. Yi *et al.*, “Add 2023: the second audio deepfake detection challenge,” *arXiv preprint arXiv:2305.13774*, 2023.
- [146] N. M. Müller, P. Czempin, F. Dieckmann, A. Froghyar, and K. Böttinger, “Does audio deepfake detection generalize?,” *arXiv preprint arXiv:2203.16263*, 2022.
- [147] H. S. Anderson and P. Roth, “Ember: an open dataset for training static pe malware machine learning models,” *arXiv preprint arXiv:1804.04637*, 2018.
- [148] A. Coates, A. Ng, and H. Lee, “An analysis of single-layer networks in unsupervised feature learning,” in *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, JMLR Workshop and Conference Proceedings, 2011, pp. 215–223.

- [149] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.
- [150] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” *arXiv preprint arXiv:1409.1556*, 2014.
- [151] C. Szegedy *et al.*, “Going deeper with convolutions,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 1–9.
- [152] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, “Mobilenetv2: Inverted residuals and linear bottlenecks,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 4510–4520.
- [153] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, “Densely connected convolutional networks,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4700–4708.
- [154] A. Conneau, D. Kiela, H. Schwenk, L. Barrault, and A. Bordes, “Supervised learning of universal sentence representations from natural language inference data,” *arXiv preprint arXiv:1705.02364*, 2017.
- [155] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “Bert: Pre-training of deep bidirectional transformers for language understanding,” in *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, 2019, pp. 4171–4186.
- [156] D. Guan and W. Zhao, “Adversarial detection based on inner-class adjusted cosine similarity,” *Applied Sciences*, vol. 12, no. 19, p. 9406, 2022.
- [157] “Catalogue of Tools & Metrics for Trustworthy AI - OECD.AI.” Accessed: Aug. 09, 2025. [Online]. Available: <https://oecd.ai/en/catalogue/tools>
- [158] W. Yue, B. Liu, and P. Stone, “t-DGR: A trajectory-based deep generative replay method for continual learning in decision making,” *arXiv preprint arXiv:2401.02576*, 2024.
- [159] S. Sarkar, A. Bansal, U. Mahbub, and R. Chellappa, “UPSET and ANGRI: Breaking high performance image classifiers,” *arXiv preprint arXiv:1707.01159*, 2017.
- [160] S. Shan, A. N. Bhagoji, H. Zheng, and B. Y. Zhao, “Poison forensics: Traceback of data poisoning attacks in neural networks,” in *31st USENIX Security Symposium (USENIX Security 22)*, 2022, pp. 3575–3592.

- [161] J.-T. Yang, H. Jiang, H. Li, D.-S. Ye, and W. Jiang, “FAD: Fine-Grained Adversarial Detection by Perturbation Intensity Classification,” *Entropy*, vol. 25, no. 2, p. 335, 2023.
- [162] J. R. Correia-Silva, R. F. Berriel, C. Badue, A. F. De Souza, and T. Oliveira-Santos, “Copycat CNN: Are random non-labeled data enough to steal knowledge from black-box models?,” *Pattern Recognit*, vol. 113, p. 107830, 2021.
- [163] C. Dunn, N. Moustafa, and B. Turnbull, “Robustness evaluations of sustainable machine learning models against data poisoning attacks in the internet of things,” *Sustainability*, vol. 12, no. 16, p. 6434, 2020.
- [164] A. M. Sadeghzadeh, S. Shiravi, and R. Jalili, “Adversarial network traffic: Towards evaluating the robustness of deep-learning-based network traffic classification,” *IEEE Transactions on Network and Service Management*, vol. 18, no. 2, pp. 1962–1976, 2021.
- [165] B. Biggio, G. Fumera, and F. Roli, “Security evaluation of pattern classifiers under attack,” *IEEE Trans Knowl Data Eng*, vol. 26, no. 4, pp. 984–996, 2013.
- [166] E. Nowroozi, Y. Mekdad, M. H. Berenjestanaki, M. Conti, and A. El Fergougui, “Demystifying the transferability of adversarial attacks in computer networks,” *IEEE Transactions on Network and Service Management*, vol. 19, no. 3, pp. 3387–3400, 2022.