Paracha *et al.*
EURASIP Journal on Information Security (2024) 2024:10
https://doi.org/10.1186/s13635-024-00158-3

EURASIP Journal on Information Security

REVIEW

Open Access

Machine learning security and privacy: a review of threats and countermeasures

Anum Paracha[1*], Junaid Arshad[1], Mohamed Ben Farah[1] and Khalid Ismail[1]

Abstract

Machine learning has become prevalent in transforming diverse aspects of our daily lives through intelligent digital solutions. Advanced disease diagnosis, autonomous vehicular systems, and automated threat detection and triage are some prominent use cases. Furthermore, the increasing use of machine learning in critical national infrastructures such as smart grids, transport, and natural resources makes it an attractive target for adversaries. The threat to machine learning systems is aggravated due to the ability of mal-actors to reverse engineer publicly available models, gaining insight into the algorithms underpinning these models. Focusing on the threat landscape for machine learning systems, we have conducted an in-depth analysis to critically examine the security and privacy threats to machine learning and the factors involved in developing these adversarial attacks. Our analysis highlighted that feature engineering, model architecture, and targeted system knowledge are crucial aspects in formulating these attacks. Furthermore, one successful attack can lead to other attacks; for instance, poisoning attacks can lead to membership inference and backdoor attacks. We have also reviewed the literature concerning methods and techniques to mitigate these threats whilst identifying their limitations including data sanitization, adversarial training, and differential privacy. Cleaning and sanitizing datasets may lead to other challenges, including underfitting and affecting model performance, whereas differential privacy does not completely preserve model's privacy. Leveraging the analysis of attack surfaces and mitigation techniques, we identify potential research directions to improve the trustworthiness of machine learning systems.

Keywords Adversarial attacks, Scrutiny-by-design, Poisoned dataset, Exploiting integrity, Data sanitization, Differential privacy

1 Introduction

Machine learning underpins significant advancements in the digital era by automating systems and making solutions autonomous and self-learned [1, 2]. Examples include facial recognition systems [3, 4], spam-filtering systems [5, 6], securing autonomous vehicle and IoT systems [7–9], and intelligent firewalls [10, 11] which puts forward the need for its security evaluation and robustness against adversarial machine learning (AML) attacks. Adversaries considerably manipulate machine learning to degrade the victim's performance, inject a backdoor, or exploit its privacy, specifically targeting security-sensitive applications [12] to disrupt their integrity or secrecy.

Breaching integrity by manipulating training datasets or model parameters is a poisoning attack. Some existing poisoning attacks are feature collision attacks [13], convex polytope attacks [14], random label flipping attacks [15, 16], and fast gradient sign method (FGSM) attack [17]. Manipulating the testing dataset is an evasion attack [18, 19]. Simultaneously, the privacy of the ML models can be exploited with model inversion or inference attacks to either reveal the parameters of the targeted model or extrapolate manipulated data to infer the expected output to analyze and assess the

*Correspondence:
Anum Paracha
anum.paracha@mail.bcu.ac.uk
[1] School of Computing and Digital Technology, Birmingham City University, Birmingham, UK

functional capabilities of the model. Bhagoji, A.N. [20] exploited the privacy of the content moderation classifier hosted by Clarifai.

Recent successful attacks on real-time machine learning systems prove the practicality of adversarial ML attacks. Zou A. et al. [21] attacked ChatGPT, Claude, and Bard with inference accuracy of 50% on GPT-4 and 86.6% on GPT-3.5. Also, Gong X. et al. [22] attacked commercial Alibaba API with a 97% success rate. These attacks highlight the urge for comprehensive research to make ML models resilient, specifically focusing on security-by-design solutions that should focus on the security and resilience of the development process rather than particular models.

In this literature review, we have performed a comparative analysis of various adversarial attack types that threaten machine learning model development: poisoning, evasion, model inversion, and membership inference attacks. Comprehensively analyzing the severity, impact, and limitations of each attack type provides valuable insights that help shape potential research directions and address the research questions given below:

- What are the significant ML adversarial attack types and attack surfaces to study and analyze?
- What is the impact of integral entities, including adversary and targeted domain, in devising adversarial attacks to exploit the victim model/system?
- What are the existing and most effective strategies and solutions to examine to mitigate ML adversarial attacks and their limitations concerning ML adversarial attack types?
- What are the open challenges and vulnerabilities identified in ML models, existing mitigation solutions, and open attack surfaces in machine learning?

Correlation between our problem statement, research questions, and interlinked addressed sections are given in Fig. 1. Addressing the above-given research questions, the major contributions that this study makes are listed as follows:

- To the best of our knowledge, this is the pioneering study to conduct an in-depth and critical analysis of threats to the machine learning systems by analyzing adversarial machine learning attacks, their severity, impact, and existing mitigation strategies and their limitations.
- Unique analysis criterion is developed to examine existing research to determine the efficacy of adversarial attack types that can be implemented on machine learning. Our criterion analyzes various attack vectors based on adversary's capability and accessibility, victim model, and technical examination of threat on machine learning model.
- Four major adversarial ML attack types are studied, based on the modeling process of machine learning, in this literature review to conduct detailed examination of attack vectors and attack surfaces in machine learning.
- Through deep analysis of threat landscape for machine learning systems and mitigation techniques proposed in existing literature, open challenges and future research directions have been identified to motivate further research and development in this area.

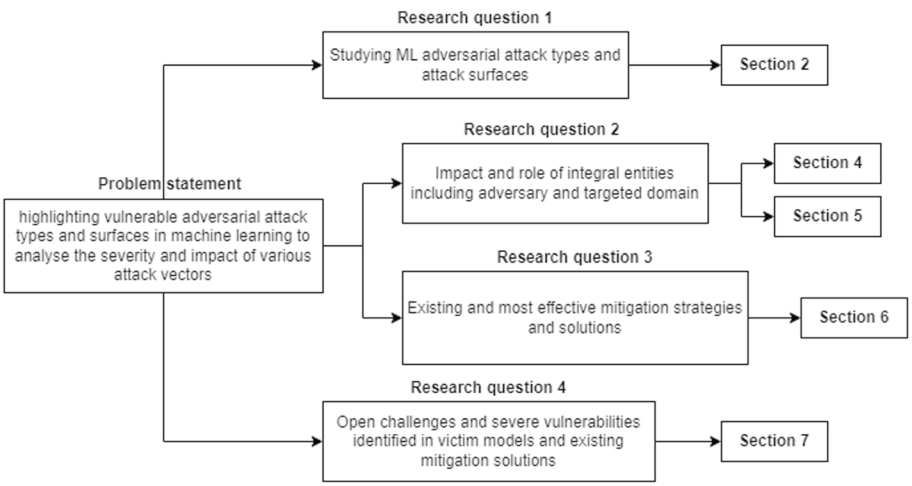

Fig. 1 Structure of this literature review to address research questions

The structure of the remaining manuscript is as follows: Section 2 explains various adversarial ML attacks. Section 3 addresses existing surveys on adversarial machine learning followed by description of our analysis criteria in Section 4. Section 5 is the state-of-the-art analysis of existing research studies considering adversarial attack types, and Section 6 explored existing mitigation strategies and their limitations. Section 7 identified potential research directions, and Section 8 concludes this research study.

2 Adversarial machine learning

Machine learning is considerably used in automating digital systems [23, 24], which makes it a tempting target for adversaries to attack and potentially harm the interconnected systems. These security violations originated in a distinctive domain associated with the security of machine learning known as adversarial machine learning [25]. Adversarial machine learning deals with malicious attempts to exploit vulnerabilities in machine learning. Every adversarial attempt is classified within one of the attack types: poisoning, evasion, model inversion, or membership inference attacks. The development of an adversarial attack focuses on many other factors, including targeting significant processing phases, attack surfaces, capability, intention, knowledge of the adversary, and availability of the victim model. Based on the machine learning development process, significant attack types on machine learning are described as follows.

2.1 Adversarial attack types-based on model processing and development

2.1.1 Poisoning attack

Training a machine learning model with the pre-processed dataset is the initial development phase, which also allows adversaries to adversaries to poison it. Poisoning attacks manipulate datasets by injecting falsified samples or perturbing the existing data samples to infect the training process and mislead the classification at test time. Poisoning the dataset is possible in two formats to disrupt the labeling strategy of the victim model known as label poisoning attack [26]. Feature perturbation, leaving the integrated label as is, is known as a clean-label poisoning attack [27]. The attack surface for poisoning attacks on machine learning is highlighted in Fig. 2.

2.1.2 Evasion attack

Attacking the machine learning model at test time is called an evasion attack. This attack intends to mislead the testing data to reduce the testing accuracy of the targeted model [28]. The ultimate objective of this attack is to misconstruct the testing input to harm the test-time integrity of machine learning. Malware generative recurrent neural network (MalRNN) is a deep learning-based approach developed to trigger evasion attacks on machine learning-based malware detection systems [29]. MalRNN evades three malware detection systems that show the expedience of evasion attacks. In addition, this attack triggers the importance of reliable security solutions to mitigate vulnerabilities in machine learning

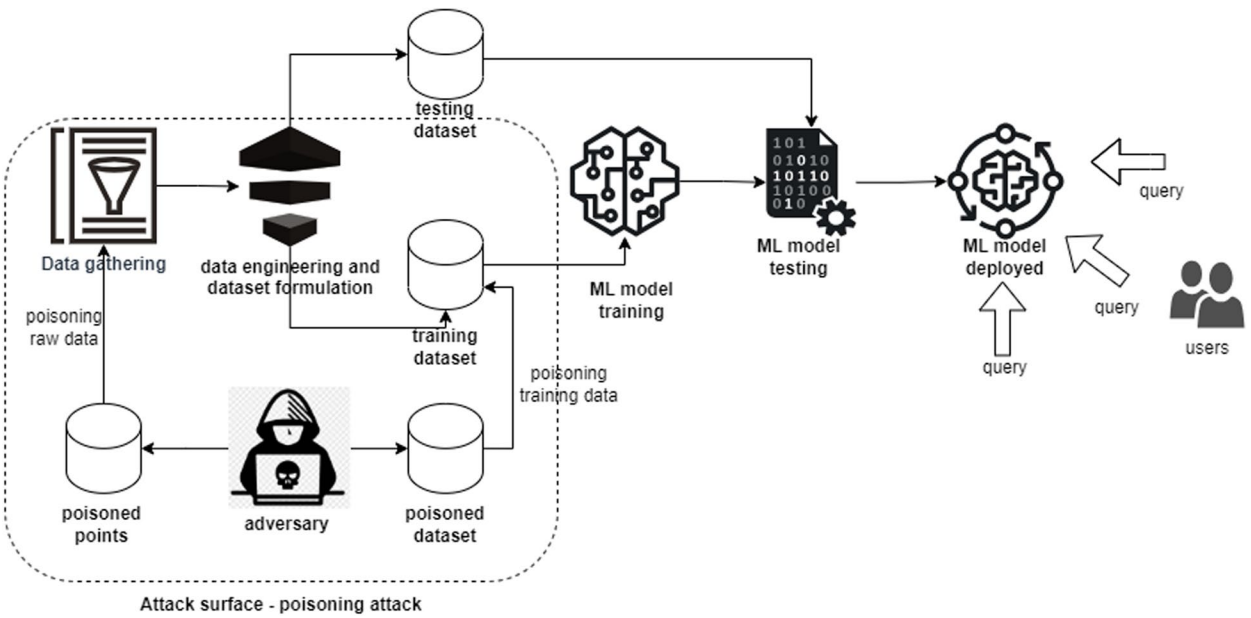

Fig. 2 Poisoning attack surface in ML model development process-poisoning ML attack

against evasion attacks. Attack surface for evasion attacks on machine learning is highlighted in Fig. 3.

2.1.3 Model inversion attack

The objective of this attack is to disrupt the privacy of machine learning. Model inversion attack is the type of attack in which an adversary tries to steal the developed ML model by replicating its underlying behavior, querying it with different datasets. An adversary extracts the baseline model representation through a model inversion attack and can regenerate the training data of the model. D. Usynin et al. [30] designed a framework for a model inversion attack on a collaborative machine learning model, demonstrating its success. It also highlights the impact of model inversion attacks on transfer machine learning models. Attack surface for model inversion attacks on machine learning is highlighted in Fig. 4.

2.1.4 Membership inference attack

A membership inference attack is another privacy attack that infers the victim model and extracts its training data, privacy settings, and model parameters. In this type of attack, the adversary has access to query the victim model under attack and can analyze the output gathered from the queried results. The adversary can regenerate the training dataset of the targeted adversarial machine learning model by analyzing the gathered queried results. The attack surface for membership inference attacks on machine learning is highlighted in Fig. 5.

Overall, machine learning model processing is at high risk of adversarial attacks. Machine learning pertains to several security and privacy vulnerabilities that exist and are exploitable at various layers of the machine learning modeling process that must be addressed adequately to mitigate adversarial attacks on machine learning models.

2.2 Adversarial attack types-based on knowledge of adversary

Adversarial attacks rely on the adversary's knowledge of the ML model under attack. When designing an adversarial attack, the adversary can have complete to zero knowledge of the target. The design of machine learning adversarial attacks is highly dependent on the knowledge of the adversary. Sub-categorizing adversarial attacks based on the adversary's knowledge is given as follows:

2.2.1 Black box attack

Black box attack is an adversarial attack for which the adversary has zero knowledge of the victim [31–33] that is put under attack. The targeted system is considered a black box for the adversary, which is the most realistic scenario because the adversary usually does not know the target system. Threat models and attack vectors are considered untargeted with the adversary's intention to reduce the overall accuracy of the targeted model. Targeted attacks can not be the scenario with the black box attack model, as the adversary does not know the victim model to exploit it with a specific targeted attack vector.

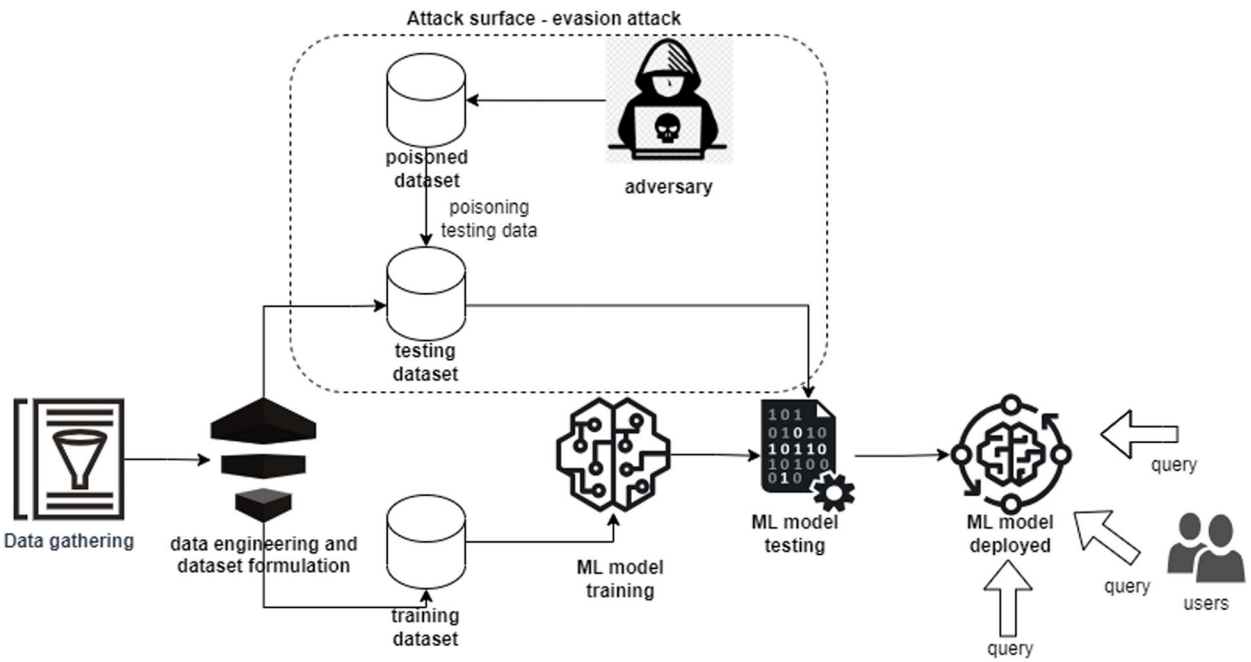

Fig. 3 Evasion attack surface in ML model development process-evasion ML attack

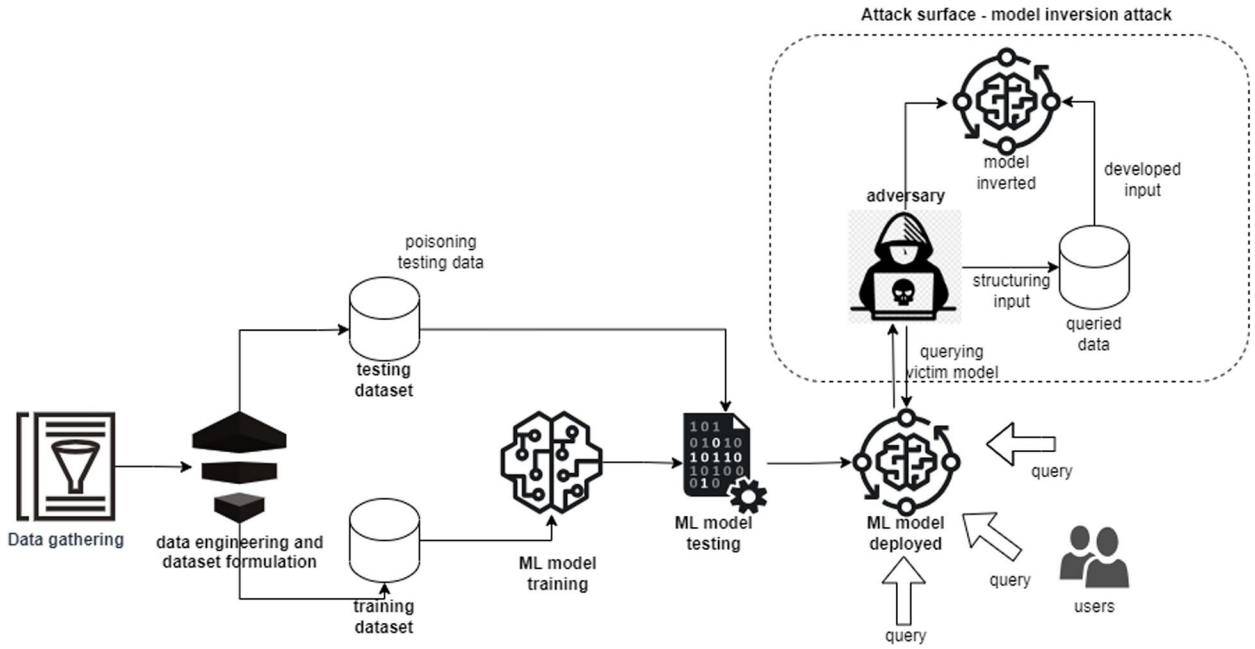

Fig. 4 Model inversion attack surface in ML model development process-model inversion ML attack

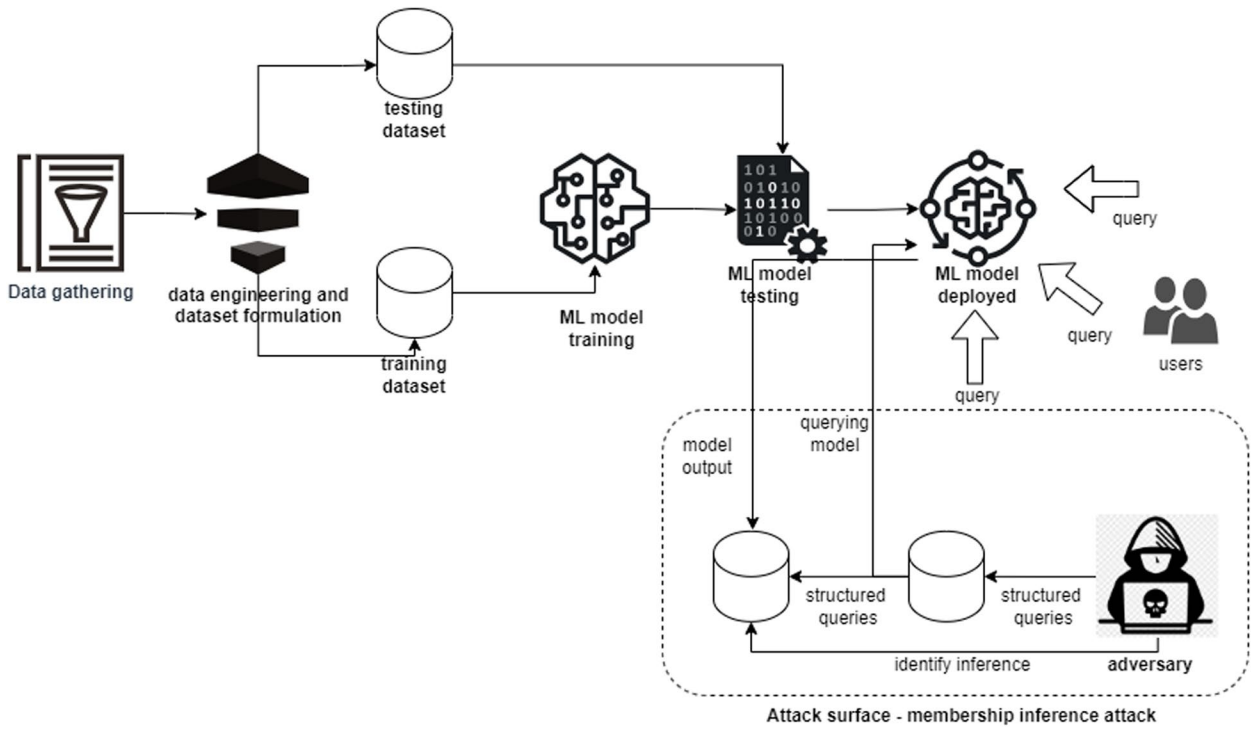

Fig. 5 Membership inference attack surface in ML model development process-membership inference ML attack

2.2.2 Gray box attack

When an adversary has partial knowledge of the target system, that kind of attack is called a gray box attack. In this case, an adversary may have some knowledge either regarding the dataset, dataset distribution, or some settings of the machine learning system that is to be

attacked [34–36]. This type of attack is more applicable to open-source systems or systems with low-security measures applied to it.

2.2.3 White box attack

White box attack is an adversarial attack where an adversary has complete knowledge of the targeted system [37–39]. This attack type is an ideal scenario where the assumption relies on the adversary having all the details of the system to be attacked. Threat models for this attack are developed considering the adversary has complete configurational knowledge of the targeted system. The white box attacks are primarily designed to achieve a specific target. These types of attacks are more applicable to poisoning and evasion attacks.

2.3 Adversarial attack types-based on capability and intention of adversary

Following the capability and intention of adversaries to attack the victim model, adversarial attacks on machine learning are additionally sub-categorized into two substantial types, highlighted below.

2.3.1 Targeted attack

Targeted attacks on machine learning systems, in adversarial settings, are formulated based on certain specified goals and targets that are the objectives of that adversarial attack [40–42]. M.K Puttagunta et al. [43] have provided a detailed synopsis of targeted and un-targeted attacks in automated medical systems. These attacks are based on the adversary's deep understanding of the targeted model and its vulnerabilities to exploit and are based on distinct aims to achieve. With this attack, the attacker has at least baseline knowledge of either the victim model or its dataset and can not be a black box attack.

2.3.2 Untargeted attack

Unlike a targeted attack, the untargeted attack is intended to disrupt the victim model in any way without any predefined objectives [44–46]. This type of attack is intended to identify the vulnerabilities of the victim machine learning model irrespective of achieving any significant goals. Generally, these attacks are black box in nature and do not explicitly define any particular data points to be used for attack, rather than the adversary intends to degrade the overall performance of the attacked ML model. Subpopulation data poisoning attack [47] is one of the case studies of untargeted adversarial attacks on machine learning.

3 Existing surveys

The research study [48] has surveyed the applicability and implication of adversarial attacks on machine learning-based cybersecurity applications, highlighting their impact. Various cyber security applications, including intrusion detection systems (IDS), biometric systems, cyber-physical systems, and spam filtering, are studied in detail from an adversarial threat perspective on how ML systems can be attacked and their integrity, confidentiality, or availability is breached, aligning knowledge and capabilities of adversaries. Also, it highlights the existing mitigation solutions against these attacks and their limitations. In conclusion, various attacks are in place that are posing threats to machine learning-based cyber security applications, which need to be addressed proactively. Although adversaries are designing more critical and complex attacks, designing complacent adversarial attacks is difficult. Overall, proactive defense systems are increasingly needed against these attacks, specifically in the cyber security domain.

M. Goldblum et al. [49] shed light on the dataset processing vulnerabilities in ML model development. Various dataset exploits and their countermeasures are reviewed in [49] to analyze the security issues in the training and processing datasets. Many training-only attacks, such as bi-level optimization, label flipping, and feature collision attacks, are highlighted, specifically on deep neural networks (DNN). Similarly, some backdoor attacks are also studied. This research paper highlights the need to mitigate dataset poisoning, which is critical and complex.

On the other hand, M. Regaki and S. Garcia [50] shed light on adversarial privacy attacks that pose another critical threat to machine learning. They studied literature from the past 7 years, with forty papers explaining the severity of privacy attacks in adversarial machine learning and their countermeasures. In this survey paper [50], authors have questioned the practical success of machine learning privacy attacks. Implementing membership inference, model extraction, reconstruction, and property inference attacks are discussed, focusing on centralized and distributed machine learning. Most attacks studied in this literature are limited to only neural networks. Some defense strategies against ML privacy attacks are also put forward, including differential privacy and regularization. But the authors thrive towards the research need in two significant directions, as given below:

- The practical implementation of the privacy attacks on real machine learning systems to prove the theoretical concepts

- Addressing limitations and enhancements of the proposed mitigation solutions to improve the privacy of machine learning models

The survey [51] provides a detailed synopsis of machine learning poisoning attacks. Questions addressed in this survey are the analysis of poisoning attack surface that leverages dataset poisoning to contaminate the training process of machine learning and model poisoning, which manipulates the machine learning model processing. Poisoning attacks studies in this literature study are the label flipping, p-tampering, and bi-level optimization on centralized machine learning, and gradient-based methods and generative approaches, including feature collision, are explored in the context of deep learning. Passive and active defense approaches are studied along with their complexities and limitations to mitigate poisoning attacks. In conclusion, limitations and further research exploration directions are identified that should be considered to address poisoning attacks in machine learning, which is still a significant and critical task to achieve.

The survey [51] provides a detailed synopsis of machine learning poisoning attacks. Questions addressed in this survey are the analysis of the dataset and model poisoning attack surfaces. Poisoning attacks in this literature study are the label flipping, p-tampering, and bi-level optimization on centralized machine learning. Gradient-based methods and generative approaches, including feature collision, are explored in deep learning. Passive and active defense approaches are studied along with their complexities and limitations to mitigate poisoning attacks. In conclusion, further research directions are identified to address poisoning attacks in machine learning (Table 1).

A. Shafee and T. A. Awaad [53] studied privacy threats in machine learning, specifically deep learning, highlighting issues and limitations in existing countermeasures. Cryptographic and perturbation techniques are studied in detail. Homomorphic encryption, functional encryption, and secure multi-party computation protocols are analyzed, determining their effect on enhancing the privacy of deep learning algorithms. Differential privacy is studied across various deep learning model layers to analyze its effectiveness in preserving privacy. The limitations of applying perturbation techniques and encryption mechanisms to secure deep learning from adversarial privacy attacks are explained. The researchers have highlighted several open research directions that help improve the privacy and confidentiality of deep learning and should not downgrade the performance and applicability of deep learning algorithms.

4 Criteria defined for literature analysis

We have conducted an in-depth literature review to analyze the complexity and criticality of existing research studies. The specific criteria developed for this detailed literature study are given in Fig. 6. Our comprehensive analysis criteria are designed based on adversarial attack types, which are further scaled down to study literature based on machine learning algorithms, datasets used to develop machine learning models, and the exploited vulnerability of machine learning algorithms. Another entity to analyze the adversarial attack and its severity is the adversary based on its knowledge and goals defined for the targeted attack on machine learning. At last, we have examined the adversarial attack severity and its impact based on the existing literature. Our developed criteria for literature analysis are given in Fig. 6. A detailed

Table 1 Comparison of related existing surveys which are peer-reviewed and focusing adversarial machine learning attacks

Research paper	Publication year	Survey type	Analysis of all attack types	Analysis criteria/ protocol	Analysis on (domain)	Solutions examined	Limitations identified
I. Rosenberg et al. [48]	2021	Traditional	✗	✗	Cyber security	✓	✓
M. Goldblum et al. [49]	2020	Traditional	✓	✗	Data poisoning	✓	✓
M. Rigaki et al. [50]	2021	Traditional	✗	✗	Privacy attacks	✓	✓
Z. Wang et al. [51]	2022	Traditional	✗	✓	Poisoning attacks	✓	✗
M. Pitropakis et al. [52]	2019	Systematic	✗	✓	Machine learning	✓	✓
A. Shafee et al. [53]	2021	Traditional	✗	✗	Privacy attacks	✓	✓
P. Bountakas et al. [54]	2023	Traditional	✗	✗	Audio, cyber-security, NLP, computer vision	✓	✓
N. Martins et al. [55]	2020	Systematic	✗	✓	Intrusion and malware detection	✓	✓
G. R. Machado et al. [56]	2021	Traditional	✗	✗	Image classification	✓	✓
A. Alotaibi et al. [57]	2023	Traditional	✗	✗	Intrusion detection system	✓	✗
This study	2023	Traditional	✓	✓	AML attack types	✓	✓

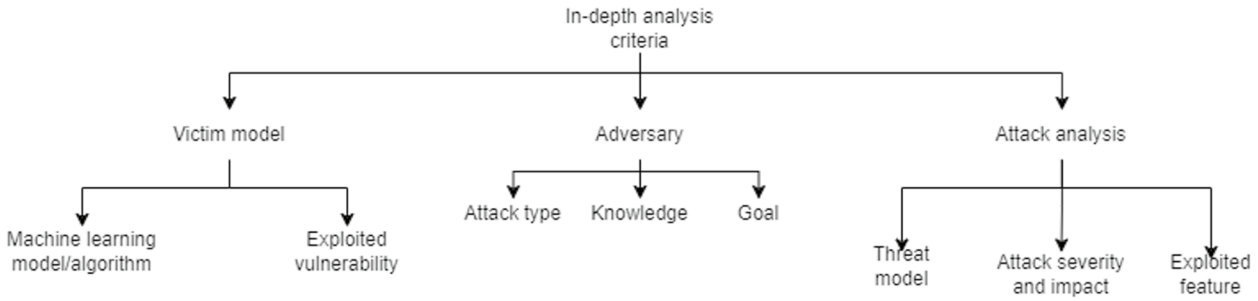

Fig. 6 Unique analysis criteria-developed to examine attacks technically w.r.t attack types

modular overview of our analysis criteria is given as follows:

- *Adversarial attack types.* The base dimension of our developed criteria is the adversarial attack types. These attack types leverage us to analyze the process-level vulnerabilities of the machine learning model. The attack types included for analysis are poisoning, evasion, model inversion, and membership inference attacks, which can be further used to devise several adversarial attacks based on these types. Analyzing existing studies based on these attack types, we have comprehensively provided a thorough summation of adaptability, implication and comparison of these attack types.
- *Machine learning model/algorithm.* The machine learning algorithm/model is also an essential aspect of our analysis as it provides the technical interpretation of the attack design under study. It is considered an influential factor in identifying the design and complexity of adversarial attacks. Also, it helps us to highlight the impact of attack type on individual machine learning algorithms.
- *Exploited vulnerability of machine learning algorithm.* Exploiting machine learning vulnerability is another essential factor in developing the attack vector to manipulate the machine learning model. This dimension helps in the technical assessment of the attack success against the targeted machine learning model. Exact annotation of the breached vulnerability helps analyze security issues in machine learning algorithms and align research directions to address these issues.
- *Knowledge level of adversary in devising adversarial attack.* Analyzing the knowledge of the adversary of the targeted model helps us better understand the attack development. Knowledge of adversary scaled from zero knowledge to completed knowledge of the targeted model or system. The adversary's knowledge is considered an important benchmark when designing these adversarial attacks. It helps analyze the impact of attacks from existing studies and compares the complexity and implication of each adversarial attack type.
- *Identifying goals of adversary.* Another significant dimension is the detailed synopsis of the adversary's goals and objectives set with the devised attack. Analyzing the intention and goals of the adversary leads to the justification of the severity of the adversarial attack. This dimension also helps technically and systematically determine security violations in the targeted model or system.
- *Attack severity and impact with respect to existing literature.* After analyzing adversarial attacks with the above mentioned dimensions, we comprehensively determine the attack severity and impact on the attacked model. Analyzing the attack severity will provide us ground to study the complexity and practical implication of adversarial attack types.

4.1 Literature review method

For providing clear and significant state-of-the-art analysis, our selection process comprises of the key concepts, defined in Section 2. In total, four reviewers have reviewed the selected papers and further refinement is particularly based on our inclusion criteria which is given as follows.

The inclusion criteria are as follows: we have selected papers that are either peer-reviewed articles or conference papers and should not go beyond 2017 as their publication year. Each paper should focused on individual adversarial attack against machine learning model and provide the technical insights of the attack development. Also, for selecting mitigation solutions papers, we have focused on the papers that provide the technical details of the developed solution and their experimental results when implemented against adversarial attack.

The exclusion criteria are as follows: for further refinement of the selected papers, we have excluded all the

papers that comprises of the comparative analysis of various adversarial attacks on machine learning model or does not provide the experimental results and insights of their developed attack and its impact on the targeted ML model.

Based on the above defined inclusion and exclusion criteria, we have developed our state-of-the-art dataset to conduct our literature analysis. For capturing inter-rater reliability of the reviewers, the Cohen's Kappa scores for each of the reviewers are 0.90, 0.93, 0.80, and 0.84, subsequently. For the detailed adversarial machine learning landscape analysis, keyword popularity is visible in Fig. 7 which highlight the impact of adversarial machine learning on various domains such as deep learning is highly inter-linked with adversarial attacks which is further affected with membership inference and model inversion attacks. Also, poisoning attacks have impacted cybersecurity, intrusion detection, and networks related applications, whereas geographical distribution of the selected papers is shown in Fig. 8 which highlight significant contributions of different countries in this domain.

4.2 Process of examining research studies

The process to examine existing literature is given in Fig. 9. This literature study has extensively examined research studies based on the adversarial attack types, studying the victim model, adversary goals, capability, attack vector, threat model, exploited features of the targeted model, and its impact. Concluding our examination process, we have provided detailed insights into the studied attack vector from a critical standpoint and in-depth forensics of the complete attack development process, highlighting and comparing the most threatening attack vectors exploiting various attack surfaces of machine learning.

This literature review examined considerable research studies based on the above-developed criteria focusing adversarial attack types.

Attacks examination, based on our analysis criteria, allows us to interpret the complete development life cycle of the adversarial attack. Studying attacks with the described dimensions reverse engineer the attack development, answering how different knowledge levels help in exploiting targeting system and, also, what features can be exploited with various attack types. At last, concluding the analysis of all the concerned entities, we have provided the impact and practicality of various adversarial machine learning attacks. Based on the criteria explained in Section 4, the filtered studies from literature are mentioned in Tables 2, 3, 4, and 5 for detailed analysis as part of this research study.

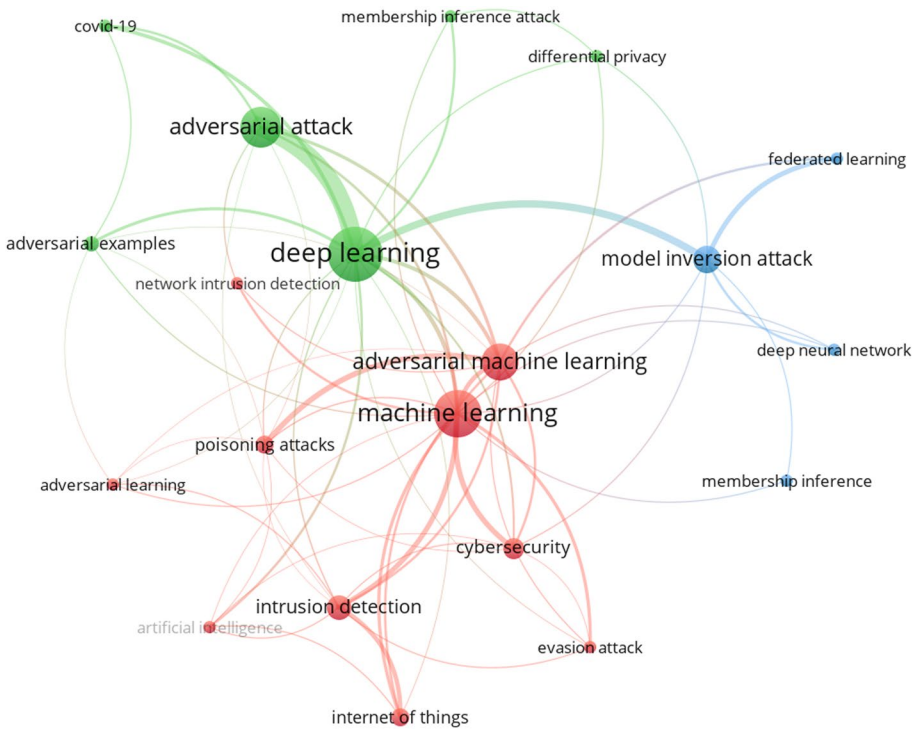

Fig. 7 Keywords analysis-a threat landscape of adversarial machine learning

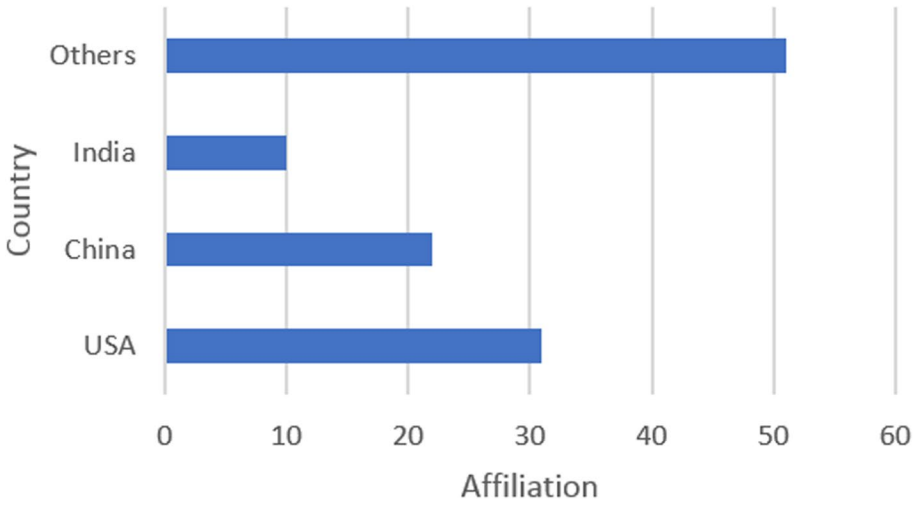

Fig. 8 Geographical distribution-an analysis of collaborative research landscape in adversarial machine learning

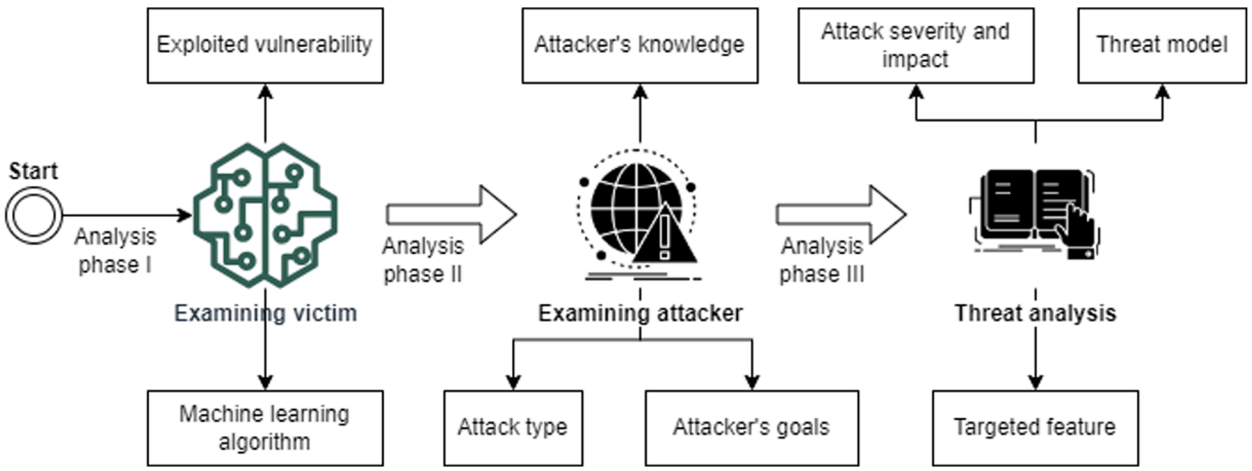

Fig. 9 Literature analysis process-examination methodology to analyze ML attacks

5 State-of-the-art analysis-AML attack types

Our study put forward an in-depth and comparative analysis based on four major adversarial attack types on machine learning. For the detailed forensics of various adversarial attack vectors, comprehensive criteria are devised to analyze each of the attack vectors and their entities in detail. Examination of each attack vector based on attack type analyzed victim threatened features, adversary, its capability and knowledge and attack vector and the severe impact of the attack vector on the victim model or algorithm. A hierarchical summary of articles studied for attack analysis is given in Fig. 10. Detailed analysis of examined attacks is given from Sections 5.1, 5.2, 5.3, and 5.4, analyzing attack vectors concerning their integrated attack type and surface.

5.1 Poisoning attack

F. A. Yerlikaya et al. [16] designed two label-flipping attacks to perturb six machine-learning algorithms with four datasets. Attacks are developed to poison binary classifiers to reduce model performance. M. Jagielski et al. [47] proposed an ML poisoning attack named subpopulation attack in which clustered attack points are injected, so their identification is difficult. It is a black box attack that manipulates data points by feature or cluster matching. Existing security techniques, training a regression model (TRIM), activation clustering, removing outliers on negative impact (RONI), and spectral signatures, are also applied to mitigate this attack but are ineffective against subpopulation data poisoning attack. In study [58], the poisoning and evasion attacks on machine learning are designed to highlight the transferability at the training and testing time of model development. It

Table 2 Analysis of poisoning attacks-based on devised examination criteria

Reference	Machine learning model/ algorithm	Attack type	Exploited vulnerability	Attacker's knowledge	Attacker's goals	Attack severity and impact	Defined threat model	Targeted feature
F. A. Yerlikaya et al. [16], 2022	SVM, SGD, logistic regression, random forest, Gaussian NB, K-NN	Random label and distance-based label flipping attacks	Poisoning dataset by changing class labels with two effective strategies	White box attack	Reduce performance (accuracy) of the system	KNN and random forest algorithms not much affected with label poisoning attacks	No	Model accuracy
M. Jagielski et al. [47], 2020	Convolution neural networks	Subpopulation attack	Poisoned cluster is integrated as subproportion of training dataset	Gray box attack	Misclassification targeted attack	Subpopulation attacks are difficult to detect and mitigate specifically in non-linear models	Yes	Test time prediction
A. Demontis et al. [58], 2019	SVM classifier, logistic, ridge, SVM-RBF	Training time poisoning attack	Reduced gradient loss with poisoned data points in transferable setting	White box, black box attacks	Violate model's integrity and availability	Poisoning attacks are more effective on models with large gradient space and high complexity	Yes	Model availability
C. Zhu et al. [14], 2019	Deep neural networks	Feature collision attack, convex polytope attack	Feature space with perturbed training samples	Gray box attack	Over fit target classifier with poisoned dataset	Turning dropout during training with poisoned data enhance transferability of poisoning attack in deep neural networks	Yes	Test time misclassification
M. Jagielski et al. [59], 2018	Linear regression	Statistically based regression points poisoning generation with flipped labels	Distinguishing legitimate and poisoned regression points with minimal gradient loss	Mean and covariance dependent gray box attack	Misclassification of the system	Residual filtrating mitigates poisoning attack on linear regression	Yes	Model accuracy

Table 3 Analysis of evasion attacks-based on devised examination criteria

Reference	Machine learning model/ algorithm	Attack type	Exploited vulnerability	Attacker's knowledge	Attacker's goals	Attack severity and impact	Defined threat model	Targeted feature
D. Gibert et al. [60], 2023	Generative adversarial networks	Query-free feature-based attack	Perturbed features in executable	Black box attack	Evade ML detector with malicious executable	ML detectors are vulnerable to be evaded with query-free attacks	No	Victim detection decision
H. Yan et al. [61], 2023	Logistic regression, SVM, NB, decision tree, RF, xgBoost, ANN, ensemble model	Label-based evasion attack	Poisoned labeled samples	Black box attack	Transfer adversarially crafted samples to evade	Transfer-based evasion attack is a serious threat to ML and DL	No	Test time precision
H. Bostani et al. [62], 2022	ML-based malware detector	n-gram based attack on malware classifier	Transform malware samples into benign with n-gram based incremental strategy	Black box attack with model query access	Misclassification of android malware detector	DNN are more affected by evading surrogate models comparing to linear SVM based classifier	Yes	Test time prediction
Md. A. Ayub et al. [28], 2020	Multi-layer perceptron network	Jacobian-based saliency map attack	Iterative approach to insert perturbation near sensitive feature of benign samples	White box attack	Misclassify malicious sample as benign in IDS	Multi-layer perceptron can be exploited with evasion attack with minimal model's knowledge	No	Test time prediction
Y. Shi et al. [63], 2017	Naïve Bayes classifier	Evasion attack with feed-forward neural networks	Feed poisoned samples with DL score under computed attack region	Exploratory black box attack	Misclassify test data samples	Controlled perturbations to labels and classification boundary may limit adversarial impact on DL	Yes	Model availability

Table 4 Analysis of model inversion attacks-based on devised examination criteria

Reference	Machine learning model/ algorithm	Attack type	Exploited vulnerability	Attacker's knowledge	Attacker's goals	Attack severity and impact	Defined threat model	Targeted feature
T. Titcombe et al. [64], 2021	Split neural networks	Model inversion attack on distributed ML	Steal intermediate/ distributed data from nodes in transfer learning	Black box attack	Invert intermediate stolen data into input format	Model inversion attacks are effective and dependent on input dataset	Yes	Model interception
M. Khosravy et al. [65], 2021	Deep neural networks	Images reconstruction with MIA	Regenerate model by intercepting private data of victim model by gathering output	Gray box attack	Inverted model and developed duplicate	ML is under serious threat of MIA attack with partial knowledge of system	No	Model privacy
Q. Zhang et al. [66], 2020	Deep neural networks	Stealing victim's model classes	Sample regeneration helps to determine private data of victim's model classes	White box attack	Developed surrogate model similar to the target	ML model can be inverted even if secured with differential privacy	Yes	Model privacy
Z. He et al. [67], 2019	Deep neural networks	Inverse-network attack strategy	Used un-trusted participant in collaborative system	Black box, white box, and query-free inversion attacks	Extract inference data with an un-trusted adversarial participant in collaborative network	Privacy preservation is challenging to achieve in split DNN	Yes	Model privacy
S. Basu et al. [68], 2019	Deep neural networks	Generative adversarial network approach	Extracted output from targeted network with generative inference details	White box attack	Extract model class/ inference details by replicating generative adversarial network	Machine learning can be inverted with generative samples	No	Model accuracy
U. Aïvodji et al. [69], 2019	Deep neural networks	Query-based generative adversarial network	Extract model details by interpreting queried outputs	Black box attack	Breach privacy of Convolutional neural networks (CNN)	Differential privacy is not much effective to mitigate MIA on machine learning	No	Model privacy

Table 5 Analysis of membership inference attacks-based on devised examination criteria

Reference	Machine learning model/ algorithm	Attack type	Exploited vulnerability	Attacker's knowledge	Attacker's goals	Attack severity and impact	Defined threat model	Targeted feature
Z. Zhu et al. [70], 2023	Multi-layer perceptron	MIA on sequential recommendation system	Surrogate and shadow models are designed to extract recommendations	Black box attack	Infer user recommendations	Inferring sequential recommendations leads to provide personalized details	Yes	Dataset inference
J. Chen et al. [71], 2021	Lasso regression, CNN	MIA with shadow model	Shadow model is used to mimic ground truth	White box attack	Retrieve confidential details of target model	Differential privacy mitigates MIA compromising accuracy of model	No	Model inference
M. Zhang et al. [72], 2021	Neural networks-based recommendation system	Inference attack to extract user-level details	Adversarial model is developed with theft users' private data	Black box attack	Retrieve private details of victim model	Popularity randomization is effective against MIA in recommender system	Yes	Model privacy
Y. Zou et al. [73], 2020	Deep neural networks	Transfer learning-based black box attack	No privacy-preserved in transfer learning model	Black box attack	Infer training model details with three formulated attacks	Transfer machine learning is at serious threat of MIA	Yes	Model inference
J. Jia et al. [74], 2019	Neural network	MIA against binary classifier	Interpret output confidence score to manipulate model details	Black box attack	Retrieve private training data of classifier	Existing solutions are subjective to dataset used in classifier	No	Dataset inference

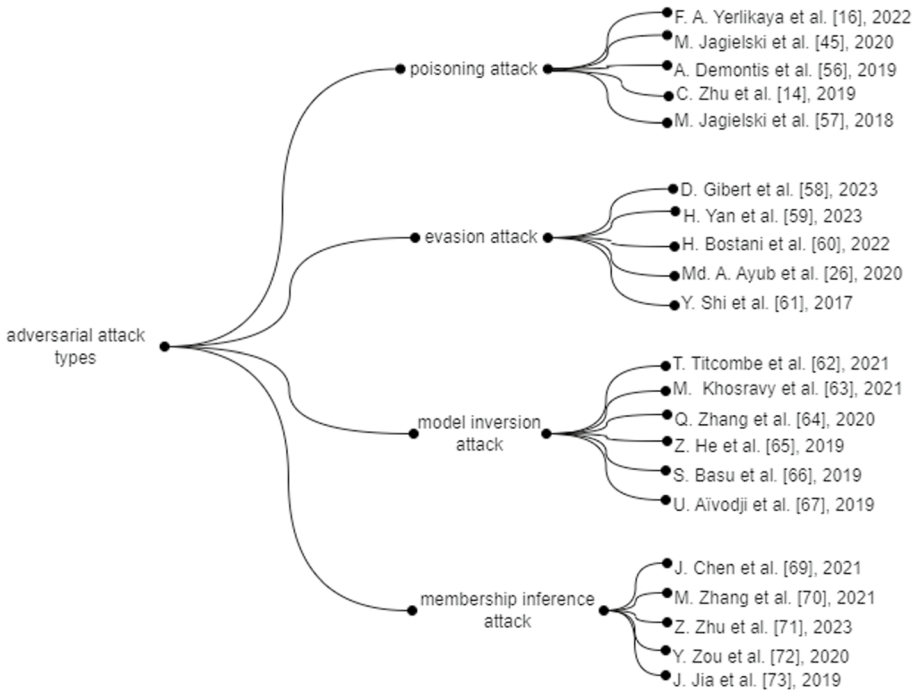

Fig. 10 State-of-the-art-AML attack types

highlights the threat of transferring poison from the surrogate to the victim model. A gradient-based optimization framework is developed to transfer poison that manipulates the gradient of input samples of training and testing datasets. This study has analyzed and highlighted the security vulnerabilities in transfer machine learning, proving empirically. This research study identifies major factors that breach integrity, making the poisoning and evasion attack successful in transfer machine learning: the attacker's optimization objectives, gradient alignment of surrogate and target model, and model complexity.

C. Zhu et al. [14] have also demonstrated the transferability of poisoning attacks in machine learning by implementing polytope attacks in deep neural networks. The impact of the poisoning attack is explained in this clean-label poisoning attack in which the adversary has poisoned only 1% of the training dataset, disrupting the results by 50%. Convex polytope attack is implemented on various deep neural networks as case studies in this research showed the sustainability and consequence of poisoning attack in transfer machine learning. This study confirmed the reliability and effectiveness of a convex polytope attack, comparing it with a feature collision attack. Also, it demonstrated the success and sustainability of transferability of convex polytope attack even in black box setting where the adversary does not know the dataset of the victim model and still achieves almost the same results as when the adversary has a 50% overlap with the target dataset. Concluding this research has formulated improvements in the transferability of poisoning attacks by turning on the dropout rate and implementing convex polytope objectives in multiple layers of neural networks. This research enforces the need to secure machine learning, specifically neural networks, from poisoning attacks in various adversarial settings.

The research study [59], particularly forensic security vulnerabilities and defense solutions of linear regression, focuses on poisoning attacks on linear regression models with gradient-based optimization and statistical attack strategies. This study has proposed a new optimization framework to poison linear regression in a gray box attack setting by evaluating the limitations of existing attacks. Another statistical-based poisoning attack is also introduced in this study, which maximizes loss by introducing poisonous points at the very corner of the boundary by exploiting the security of noise-resilient and adversarially-resilient regression. However, TRIM has been proposed, proving to be more effective in mitigating poisoning attacks in the linear regression model but ineffective against subpopulation attack, thus proving the severity of the poisoning attack in adversarial settings.

5.2 Evasion attack

Malware classifiers are also affected by adversarial attacks. In study [62], the researchers have developed a test time attack on an Android malware classifier to

disrupt its classification outcome. The attack developed in this research is a black box, which extracts the opcodes with the n-Grams strategy from the disassembled Android application packages (APK) and transforms benign samples into malicious ones with a random search technique. This attack is experimented with five different malware detectors. It proves the effectiveness of a test time attack that evades the machine learning model and misclassifies the test time classification results. As a result, machine learning-based malware detectors, including Drebin, detection malware in android (MaMaDriod), with an accuracy of 81% and 75%, respectively, and some others failed to detect malicious Android applications.

Similarly, the stealthiness of the evasion attack is explained by another attack named the Jacobian-based saliency map attack (JSMA). JSMA is developed against IDS and is designed on a multi-layer perceptron algorithm. Targeted misclassification is intended to be achieved when the adversary has intended to classify malware traffic in network intrusion detection systems (NIDS) as benign. The experimental analysis uses the white box setting to devise this evasion attack and achieved maximum accuracy drop to 29% with the TRabID 2017 dataset. Hence, it proved the malignant approach to threat machine learning-based applications in cybersecurity, subsequently highlighting the test time security vulnerabilities in neural networks.

Based on our devised criteria, we have also examined the sensitivity of evasion and causative attacks [63] against deep learning to technically shed light on the existing security vulnerabilities in deep learning that can be exploited by adversaries to harm the system. This research devised an adversarial perturbation approach and tested it with text and image datasets. At first, an evasion attack is performed, followed by the exploratory attack intended to infer the trained classification model and extract its private tuning parameters. The exploratory attack is a black box query-based attack replicating the victim model based on the obtained query outputs. With the replicated model, this attack is further intended to poison labels of testing samples and fool the deep learning model with an evasion attack.

5.3 Model inversion attack

Adversarial attacks also threaten the privacy of machine learning. Research study [67] experimentally revealed the privacy attack during inference in collaborative machine learning and argued that a single malicious participant could infer the target system and steal the confidential information of the targeted system. This attack is successful in all three settings with complete knowledge, zero knowledge, and query-free attack setting. The confidential tuning parameters are extracted, specifically with a regularized maximum likelihood estimation technique in which the adversary follows the Euclidean distance estimation and finds the optimal sample with the least variation. In conclusion, this research highlighted the potential of inference attacks that demand attention to mitigate and ensure privacy preservation of deep learning. S. Basu et al. [68] demonstrated the privacy issues in machine learning algorithms by inverting a deep neural network (DNN) with a model inversion attack. This research study implemented the model inversion attack on the facial recognition system and extracted the class representation of the model. The attack developed in this research has baseline knowledge of the target system. A generative adversarial network is integrated to generate input samples and invert the victim model, highlighting the effectiveness of generative AI in inverting the model. Another framework, named generative adversarial model inversion (GAMIN), by U. Aivodji and others [69] is also based on generative adversarial networks to craft adversarial images to query the targeted model and extract its details by comparative output resemblance. The major threat disclosed with adversarial networks is that even without prior knowledge of the system under attack, the adversary can extract its confidential settings parameters and invert it. M. Khosravy et al. [75] also developed a model inversion attack on a deep neural network-based face recognition system. It is a gray box attack as the adversary has partial knowledge of the system under attack, including the model structure and its parameters. This attack extracts the model configurations by reconstructing images based on the confidence achieved from the targeted model, hence inverting the targeted CNN model. Concluding all the mentioned attacks, the emphasis is on the privacy preservation of machine learning, which is a primary consideration in constructing trustworthy and resilient AI/ML that overcome adversarial attacks.

5.4 Membership inference attack

The membership inference attack (MIA) is another privacy risk to machine learning and deep learning. Yang Zou et al. [73] have comprehensively studied membership inference attacks on deep learning models in transfer learning mode. 95% accuracy, area under curve (AUC), is achieved with the membership inference attack performed to determine if the input instance is part of the training dataset of the targeted model. Three attacks originated in three different transfer learning modes as part of this research. When the adversary has access to the teacher model, the adversary targets the trained student model, and the adversary infers the teacher model dataset with access to the student model. This study

implemented a surrogate model based on ResNet20 convolutional neural networks with derived and student datasets and determined the membership inference of the victim model. This attack vector is quite adequate in demonstrating the capability of the inference attack on machine learning to exploit its privacy even with limited access or information of the victim model. Another potential privacy attack is mentioned in [72], where the attacker acquired an automated recommender system membership inference. The attack is declared zero-knowledge. However, this study interrogates a serious privacy threat on the recommender system's sensitive user data, which adversaries can reveal with the determined query-based attack. Here, the inference attack is defined by three recommender algorithms: item-based collaborative filtering, latent factor model, and neural collaborative filtering by implementing a shadow model to mimic the training dataset of the victim, which ultimately jeopardizes its privacy.

6 Mitigation strategies and limitations

Various mitigation techniques are also developed to secure machine learning models alongside the above-mentioned adversarial attacks. However, the existing solutions are subjective and attack-focused, which cannot guard targeted models when attacked with new techniques. Also, proposed security solutions have several limitations that should be considered to keep the integrity of machine learning intact, making AI/ML secure and trustworthy. A hierarchical description of analyzed mitigation techniques, based on adversarial attack types, is given in Fig. 11. A detailed analysis of existing security solutions based on adversarial attack types is given as follows:

6.1 Mitigating poisoning attack

6.1.1 Data sanitization

Pre-processing training datasets and removing erroneous or poisoned data points is known as data sanitization. However, this reduction may lead to a lessened dataset, increasing underfitting issues in model development. S. Venkatesan et al. [76] proposed a solution to overcome the limitations of data sanitization by creating random training data subsets to train ten ensemble classifiers to balance the poisoning effect. This mechanism reduces poisoning effects while training NIDS to 30%. Similarly, another data sanitization derivative is applied to malware detection systems to mitigate clean label poisoning

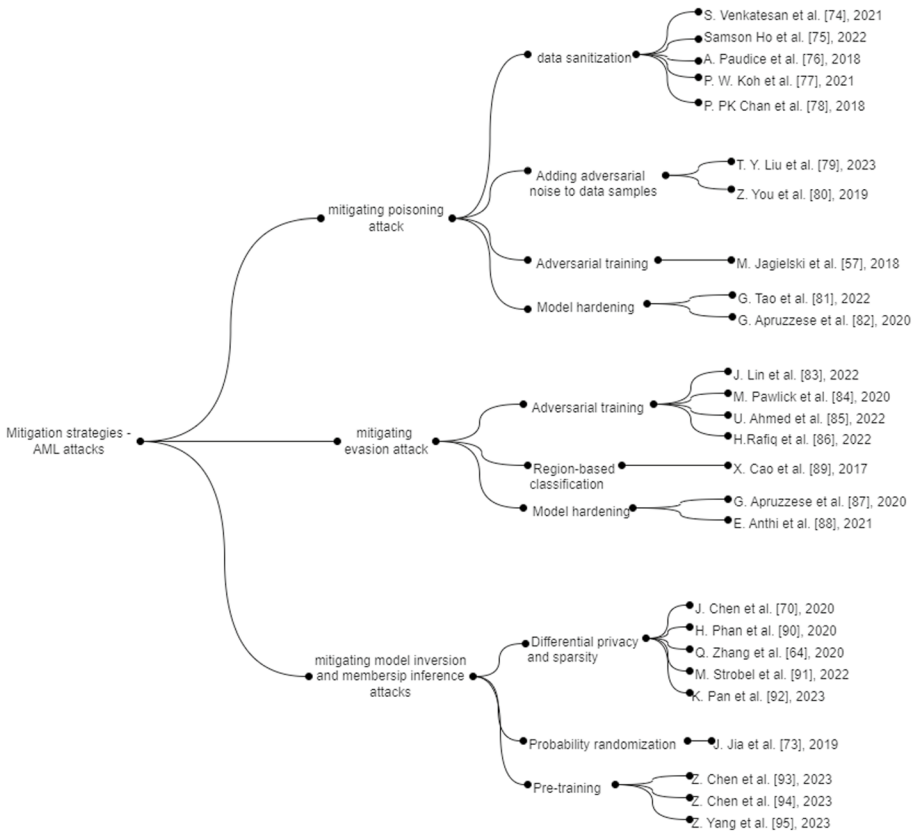

Fig. 11 State-of-the-art of mitigation techniques-AML attack types

attacks [77]. This approach is an enhancement, provided in [76]. Furthermore, the study [78] has proposed another approach to label sanitization to reduce the impact of overfitting and underfitting issues, whereas P. W. Koh and others [79] have introduced three sophisticated poisoning attacks by introducing cluster-based poisoning that breached the sanitization solutions, highlighted above.

RONI is also a derivation of data sanitization proposed by Patrick P. K. Chan and others [80], which removes poisoned data samples by analyzing the negative impact of each data sample that reduces classification accuracy. However, it also leads to underfitting issues that lessen the flexibility and increase true negatives at test time.

6.1.2 Adding adversarial noise to data samples

Training the ML model with an adversarially developed dataset allows the trained model to identify poisoned samples at test time. T. Y. Liu et al. [81] have boosted the immunity of the model by adding specifically crafted noise samples in the dataset during training, which is effective against bulls-eye polytope, gradient masking and sleeper agent attacks. Another study [82] has introduced adversarial noise into the intermediate layer of CNN to mitigate FGSM attacks.

6.1.3 Adversarial training

Training an ML model with adversarial data samples allows it to be resilient against poisoning attacks. TRIM is one of the techniques used to adversarially train models with a residual subset of a dataset with a minimum error rate. M. Jagielski and others [59] have designed and experimented with this TRIM algorithm against adversarial poisoning attacks against linear regression algorithm to solve optimization problems. This approach has reduced the error rate to approximately 6%. It performs robustly compared to random sample consensus (RANSAC), a data sanitization derivative, whereas TRIM and RONI security techniques failed against the subpopulation attack developed in [47].

6.1.4 Model hardening

Another innovative technique to mitigate poisoning attacks is model hardening, in which the model is trained until it leads to large class distances where it should not accept outliers. This technique makes it challenging for an adversary to poison the model. G. Tao et al. [83] proposed a model hardening mechanism with additional training to increase the class distances and challenge the label-flipping attack. The study [84] hardens random forest algorithm to mitigate poisoning impact on an IDS. Moreover, it also leads to mitigate backdoor attacks against neural networks. It reduces misclassification up to 80%, but it is still only effective against label-flipping backdoor attacks.

6.2 Mitigating evasion attack

6.2.1 Adversarial training

Adversarial training is a prominent mechanism to mitigate evasion attacks in machine learning. A particular dataset part is intentionally poisoned to lessen the test time evasion and make the model adversarially robust [85]. It allows the victim to be aware of adversarial samples if injected at test time to detect and defend itself if attacked by an adversary. U. Ahmed et al. [86] proposed adversarial training by classifying adversarial and normal data samples followed by centroid-based clustering of features and calculating the cosine similarity and centroid of the image vector. The research [87] to train independent models to reduce fabricated classification attacks and [88] guards against Carlini and Wagner and FGSM attacks.

6.2.2 Model hardening

The hardening machine learning model also applies to developing a wall of security in machine learning against adversarial attacks at test time. Evasion attacks are also mitigated with the help of a training model until they reach the state of hardening, which activates the model to evade adversaries and mitigate attack impact. Adversarially crafted samples are injected intentionally during the machine learning model training to evade the system until it reaches the state of hardening, making the victim model resilient and robust. These poisoned input data samples evade the system and are then marked as poisoned in the system to identify similar patterns if injected by the adversary at test time. G. Apruzzese et al. [89] has introduced a similar strategy to mitigate evasion attacks in botnet detection systems by deep reinforcement learning. They have developed an agent based on deep reinforcement learning capable of generating adversarial samples to evade the targeted botnet and then including these adversarially generated samples into the targeted system marked as malicious to make the model understand the pattern of adversarial samples if attacked during test time, whereas research study [90] used model hardening to secure ML-based IoT system. A threshold is specified that trains the model properly with the legitimate and illegitimate dataset that makes the botnet detector robust against evasion attacks.

6.2.3 Region-based classification

X. Cao et al. [91] have designed a classification mechanism based on region rather than individual sample points. The researchers provided this technique based on the assumption that the adversarial points lie near the

classification boundary. A hypercube-centered classification approach is determined by omitting single-point-based classification at test time to reduce the impact of adversarial points.

6.3 Mitigating model inversion and membership inference attacks

6.3.1 Differential privacy and sparsity

To preserve the privacy of machine learning models, one of the profound solutions is differential privacy. It makes it difficult for the adversary to analyze the output and extract the victim's confidential information. J. Chen et al. [71] have used differential privacy applied with stochastic gradient descent on Lasso and CNN neural to preserve genomic data privacy. H. Phan et al. [92] improves DNN robustness by implementing differential privacy with the logarithmic relation between the privacy budget and the accuracy of the targeted model. They have empirically analyzed genomic data for phenotype prediction with a white box attack, whereas Q. Zhang et al. [66] improves differential privacy by implementing it at class and sub-class level, proving the minimal probability of model inversion attack at dataset only. Class and sub-class level differential privacy is more effective and robust than simple record-level differential privacy, providing more Euclidean distance between original and inverted data samples. However, it is tested with neural networks only with Face24 and MNIST datasets. Also, this type of differential privacy requires high computational resources, whereas the study [93] highlights trade-offs of data privacy and assuring its trustworthiness. K. Pan et al. [94] implemented differential privacy to mitigate privacy attacks and data leaks against generative adversarial networks, whereas the floating-point attack mentioned in [95] has invalidated differential privacy implemented to preserve privacy of machine learning models.

6.3.2 Probability randomization

Adversarial privacy attacks, specifically membership inference attacks, target machine learning classifiers and infer input datasets by interpreting the confidence score and probability of the queried output. Adding noise to the output or intentionally interrupting the confidence probability score leads to the privacy preservation of machine learning, preventing adversaries from inferring confidential details of the victim model. Membership inference guard (MemGuard) [74] is one of the solutions designed to preserve the privacy of machine learning models against membership inference attacks by adding randomized noise to each of the score vectors with a specified probability of accuracy loss and makes machine learning-based binary classifier resilient to mitigate membership inference attack. However, the solution is only tested for securing neural networks under the black box attack settings.

6.3.3 Pre-training

Z. Chen et al. [96] have proposed a model-preserving framework to preserve the security of deep learning models while training models by combining model parameters and training data. Z. Chen et al. [97] have introduced a new framework to pre-train an ML-based model to preserve privacy by enforcing less confidence in the queried results between members and non-members. Z. Yang and others [98] have introduced another model to statistically in-distinguish the confidence scores of members and non-members.

7 Potential research directions-AML attack types

Machine learning is at the edge of adversarial attacks that threaten the security and privacy of machine learning. In this literature review, we have analyzed the existing adversarial attacks on machine learning, their mitigation strategies, and limitations based on adversarial machine learning attack types. Based on these attack types, the following are the potential research directions that can be extended as future research.

- To make machine learning safe and resilient against security attacks that disrupt its integrity, we need to improve mitigation solutions and develop solutions that make machine learning secure by design and robustify its model development process. Some prominent solutions are highlighted in existing literature but are subjective to mitigate vulnerabilities of a particular system to secure that specific domain or system environment but can not fight any new attack if implemented.
- Another important perspective towards trustworthy machine learning is preserving its privacy against model inversion attacks [30, 75, 99, 100] and membership inference attacks [101–103] that violates its secrecy. Both of these attacks need to be addressed to preserve the privacy of machine learning and make machine learning explainable and reliable to use.
- Identifying the reliability, integrity, and usability of machine learning in security-sensitive applications such as financial recognition systems [104–106], medical diagnostic applications [107, 108], medical imaging systems [43, 109, 110], and cyber defenses [111–113] is potentially a critical and open research challenge. The significance and prevalence of secure and trustworthy machine learning are prominently highlighted by its use in these domains. At the same time, persisting threats and existing vulnerabilities

should be of greater concern to be resolved significantly to ensure the reliable use of machine learning.
- Another important research direction is the practical implications of theoretical adversarial attacks on machine learning. Many research studies and surveys claimed that most of the adversarial attacks are highlighted in a theoretical manner [114] and maybe just implemented as white box attacks, which are less credible practically. The practical implication of these attacks and defenses is an open research challenge that should be particularly considered to highlight the impact of adversarial attacks in reality.

Overall, many security and privacy-preserving solutions are provided in the literature. Still, to the best of our knowledge, security solutions and strategies given in the literature are very subjective in nature and target specific attack vectors with limited datasets in particular domains or systems to be implemented. Context-aware solutions against these adversarial attacks on machine learning are a potential research challenge that should be focused on.

8 Conclusion

We have conducted a comprehensive study to analyze different types of adversarial attacks, their development process, and their impact alongside defenses and limitations. For the in-depth analysis, various aspects of malicious attempts are studied, including the adversary's knowledge and accessibility, adaptations to algorithms, vulnerability, and feature exploitation. Existing defense mechanisms are also studied to mitigate adversarial attacks, including data sanitization, outlier detection, adversarial training, and differential privacy and sparsity. Moreover, their limitations and successful attacks that breached these security techniques are highlighted to provide a structured ground and deeper insights for further investigations. Our study provides a detailed comparative analysis of adversarial attack types, investigating the significance of various technical aspects and providing deeper insights into their development process.

Our analysis highlights the ability of adversaries to develop adversarial attacks to breach machine learning security and privacy. Poisoning attacks are identified as a major threat to machine learning, whereas practical implications of inference attacks, such as attacks developed in [21] against large language models, highlighted their impact. We have concluded that the public availability of the datasets and models gives provenance to the adversaries to exploit ML models even with zero knowledge of the targeted models. Also, adversarial attacks are transferable, allowing adversaries to penetrate the targeted model with the help of surrogate models. For example, the attack developed in [14] is transferable. Also, it is important to consider the security of the machine learning development process while developing mitigation solutions to counter adversarial attacks.

Authors' contributions
Anum Paracha-she is responsible for the development of the idea, conducting the literature analysis and writing the manuscript. Junaid Arshad-he is responsible for the refinement of the idea and supervision of this literature analysis. Also, he has reviewed and upgrade the writing of this manuscript. Mohamed Ben Farah-he has contributed in reviewing the article. Khalid Ismail-he has contributed in reviewing the article.

Funding
This work was supported by the U.K. West Midlands Innovation Accelerator WMHTIA Grant 10056871.

Availability of data and materials
The compiled and refined data and results of this manuscript should be available upon request by the authors.

Declarations

Ethics approval and consent to participate
This literature analysis is solely based on the gathered data from the Scopus for which no additional permissions are required.

Competing interests
The authors declare no competing interests.

Received: 19 November 2023 Accepted: 2 April 2024
Published online: 23 April 2024

References

1. R. Rosati, L. Romeo, G. Cecchini, F. Tonetto, P. Viti, A. Mancini, E. Frontoni, From knowledge-based to big data analytic model: a novel iot and machine learning based decision support system for predictive maintenance in industry 4.0. J. Intell. Manuf. **34**(1), 107–121 (2023)
2. B. Jothi, M. Pushpalatha, Wils-trs-a novel optimized deep learning based intrusion detection framework for iot networks. Pers. Ubiquit. Comput. **27**(3), 1285–1301 (2023)
3. A. Singh, S. Bhatt, V. Nayak, M. Shah, Automation of surveillance systems using deep learning and facial recognition. Int. J. Syst. Assur. Eng. Manag. **14**(Suppl 1), 236–245 (2023)
4. S. Gupta, P. Kumar, R.K. Tekchandani, Facial emotion recognition based real-time learner engagement detection system in online learning context using deep learning models. Multimedia Tools Appl. **82**(8), 11365–11394 (2023)
5. D. Komarasamy, O. Duraisamy, M.S. S, S. Krishnamoorthy, S. Rajendran, D.M. K, in *2023 7th International Conference on Computing Methodologies and Communication (ICCMC)*, Spam email filtering using machine learning algorithm. IEEE. 1–5 (2023). https://www.ieee.org/conferences/publishing/index.html.
6. W.M. Salama, M.H. Aly, Y. Abouelseoud, Deep learning-based spam image filtering. Alex. Eng. J. **68**, 461–468 (2023)
7. C. Chen, C. Wang, B. Liu, C. He, L. Cong, S. Wan, Edge intelligence empowered vehicle detection and image segmentation for autonomous vehicles. IEEE Trans. Intell. Transp. Syst. **24**, 13023–13034 (2023)
8. S. Feng, H. Sun, X. Yan, H. Zhu, Z. Zou, S. Shen, H.X. Liu, Dense reinforcement learning for safety validation of autonomous vehicles. Nature **615**(7953), 620–627 (2023)
9. S. Menon, D. Anand, Kavita, S. Verma, M. Kaur, N. Jhanjhi, R.M. Ghoniem, S.K. Ray, Blockchain and machine learning inspired secure smart home communication network. Sensors **23**(13), 6132 (2023)

10. M.H. Rahman, T. Islam, M.M. Rana, R. Tasnim, T.R. Mona, M.M. Sakib, Machine Learning Approach on Multiclass Classification of Internet Firewall Log Files. In 2023 International Conference on Computational Intelligence and Sustainable Engineering Solutions (CISES). 358-364 (2023). IEEE. publisher address: https://www.ieee.org/conferences/publishing/index.html
11. K. Surendhar, B.K. Pandey, G. Geetha, H. Gohel, in *2023 IEEE 12th International Conference on Communication Systems and Network Technologies (CSNT)*. Detection of payload injection in firewall using machine learning (IEEE, 2023), pp. 186–190 https://www.ieee.org/conferences/publishing/index.html
12. O. Oyebode, J. Fowles, D. Steeves, R. Orji, Machine learning techniques in adaptive and personalized systems for health and wellness. Int. J. Hum. Comput. Interact. **39**(9), 1938–1962 (2023)
13. A. Shafahi, W.R. Huang, M. Najibi, O. Suciu, C. Studer, T. Dumitras, T. Goldstein, Poison frogs! targeted clean-label poisoning attacks on neural networks. Adv. Neural Inf. Process. Syst. **31** 6103–6113 (2018)
14. C. Zhu, W.R. Huang, A. Shafahi, H. Li, G. Taylor, C. Studer, T. Goldstein, Transferable clean-label poisoning attacks on deep neural nets. In International conference on machine learning. 7614-7623 2019. PMLR. International Conference on Machine Learning (ICML). https://www.proceedings.com/international-conference-on-machine-learning-icml/
15. M.A. Ramirez, S. Yoon, E. Damiani, H.A. Hamadi, C.A. Ardagna, N. Bena, Y.J. Byon, T.Y. Kim, C.S. Cho, C.Y. Yeun, New data poison attacks on machine learning classifiers for mobile exfiltration. arXiv preprint. arXiv:2210.11592. 2022
16. F.A. Yerlikaya, Şerif Bahtiyar, Data poisoning attacks against machine learning algorithms. Expert Syst. Appl. **208**, 118101 (2022)
17. B. Pal, D. Gupta, M. Rashed-Al-Mahfuz, S.A. Alyami, M.A. Moni, Vulnerability in deep transfer learning models to adversarial fast gradient sign attack for covid-19 prediction from chest radiography images. Appl. Sci. **11**(9), 4233 (2021)
18. T. Combey, A. Loison, M. Faucher, H. Hajri, Probabilistic jacobian-based saliency maps attacks. Mach. Learn. Knowl. Extraction **2**(4), 558–578 (2020)
19. R. Wiyatno, A. Xu, Maximal jacobian-based saliency map attack. arXiv preprint arXiv:1808.07945. 2018
20. A.N. Bhagoji, W. He, B. Li, D. Song, Exploring the space of black-box attacks on deep neural networks. arXiv preprint arXiv:1712.09491. 2017
21. A. Zou, Z. Wang, N. Carlini, M. Nasr, J.Z. Kolter, M, Fredrikson Universal and transferable adversarial attacks on aligned language models. arXiv preprint arXiv:2307.15043. 2023
22. X. Gong, Y. Chen, W. Yang, H. Huang, Q. Wang, b3: Backdoor attacks against black-box machine learning models. ACM Trans. Privacy Secur. **26**, 1–24 (2023)
23. A. Awajan, A novel deep learning-based intrusion detection system for iot networks. Computers **12**(2), 34 (2023)
24. H. Shah, D. Shah, N.K. Jadav, R. Gupta, S. Tanwar, O. Alfarraj, A. Tolba, M.S. Raboaca, V. Marina, Deep learning-based malicious smart contract and intrusion detection system for iot environment. Mathematics **11**(2), 418 (2023)
25. D. Rios Insua, R. Naveiro, V. Gallego, J. Poulos, Adversarial machine learning: Bayesian perspectives. J. Am. Stat. Assoc. **118**, 1–12 (2023)
26. P. Gupta, K. Yadav, B.B. Gupta, M. Alazab, T.R. Gadekallu, A novel data poisoning attack in federated learning based on inverted loss function. Comput. Secur. **130**, 103270 (2023)
27. B. Zhao, Y. Lao, in *Proceedings of the AAAI Conference on Artificial Intelligence*. Clpa: Clean-label poisoning availability attacks using generative adversarial nets. The AAAI Press. **36**, 9162–9170 (2022). https://aaai.org/aaai-publications/
28. M.A. Ayub, W.A. Johnson, D.A. Talbert, A. Siraj, in *2020 54th Annual Conference on Information Sciences and Systems (CISS)*. Model evasion attack on intrusion detection systems using adversarial machine learning. IEEE. 1–6 (2020) https://www.ieee.org/conferences/publishing/index.html
29. M. Ebrahimi, N. Zhang, J. Hu, M.T. Raza, H. Chen, Binary Black-box Evasion Attacks Against Deep Learning-based Static Malware Detectors with Adversarial Byte-Level Language Model. In 2021 AAAI workshop on Robust, Secure and Efficient Machine Learning (RSEML). The AAAI Press. 2021. https://aaai.org/conference/aaai/aaai21/ws21workshops/
30. D. Usynin, D. Rueckert, G. Kaissis, Beyond gradients: Exploiting adversarial priors in model inversion attacks. ACM Trans. Priv. Secur. **26**,(3) 1–30 (2023)
31. Y. Bai, Y. Wang, Y. Zeng, Y. Jiang, S.T. Xia, Query efficient black-box adversarial attack on deep neural networks. Pattern Recog. **133**, 109037 (2023)
32. M. Yu, S. Sun, in *Proceedings of the AAAI Conference on Artificial Intelligence*. Natural black-box adversarial examples against deep reinforcement learning. The AAAI Press. **36**, 8936–8944 (2022). https://aaai.org/aaai-publications/
33. C. Sun, Y. Zhang, W. Chaoqun, Q. Wang, Y. Li, T. Liu, B. Han, X. Tian, Towards lightweight black-box attack against deep neural networks. Adv. Neural Inf. Process. Syst. **35**, 19319–19331 (2022)
34. H. Wang, S. Wang, Z. Jin, Y. Wang, C. Chen, M. Tistarelli, in *2021 16th IEEE international conference on automatic face and gesture recognition (FG 2021)*. Similarity-based gray-box adversarial attack against deep face recognition (IEEE, 2021), pp. 1–8 https://www.ieee.org/conferences/publishing/index.html
35. N. Aafaq, N. Akhtar, W. Liu, M. Shah, A. Mian, Language model agnostic gray-box adversarial attack on image captioning. IEEE Trans. Inf. Forensic Secur. **18**, 626–638 (2022)
36. R. Lapid, M. Sipper, 2023. I See Dead People: Gray-box adversarial attack on image-to-text models. arXiv preprint arXiv:2306.07591.
37. W. Patterson, I. Fernandez, S. Neupane, M. Parmar, S. Mittal, S. Rahimi, A white-box adversarial attack against a digital twin. (2022). arXiv preprint arXiv:2210.14018
38. S. Agnihotri, S. Jung, M. Keuper, CosPGD: A unified white-box adversarial attack for pixel-wise prediction tasks. 2023. arXiv preprint arXiv:2302.02213
39. D. Wu, S. Qi, Y. Qi, Q. Li, B. Cai, Q. Guo, J. Cheng, Understanding and defending against white-box membership inference attack in deep learning. Knowl. Based Syst. **259**, 110014 (2023)
40. A. Guesmi, K.N. Khasawneh, N. Abu-Ghazaleh, I. Alouani, in *2022 International Joint Conference on Neural Networks (IJCNN)*. Room: Adversarial machine learning attacks under real-time constraints (IEEE, 2022), pp. 1–10 https://www.ieee.org/conferences/publishing/index.html
41. E. Abdukhamidov, M. Abuhamad, G.K. Thiruvathukal, H. Kim, T. Abuhmed, Single-Class Target-Specific Attack against Interpretable Deep Learning Systems. (2023). arXiv preprint arXiv:2307.06484.
42. W. Feng, N. Xu, T. Zhang, Y. Zhang, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Dynamic generative targeted attacks with pattern injection. IEEE. 16404–16414 (2023). https://www.ieee.org/conferences/publishing/index.html
43. M.K. Puttagunta, S. Ravi, C. Nelson Kennedy Babu, Adversarial examples: attacks and defences on medical deep learning systems. Multimedia Tools Appl. **82**, 1–37 (2023)
44. A. Zafar, et al., Untargeted white-box adversarial attack to break into deep leaning based covid-19 monitoring face mask detection system. Multimedia Tools Appl. **83**, 1–27 (2023)
45. B. Chen, Y. Feng, T. Dai, J. Bai, Y. Jiang, S.T. Xia, X. Wang, Adversarial examples generation for deep product quantization networks on image retrieval. IEEE Trans. Pattern Anal. Mach. Intell. **45**(2), 1388–1404 (2023)
46. Y. Li, Z. Li, L. Zeng, S. Long, F. Huang, K. Ren, Compound adversarial examples in deep neural networks. Inf. Sci. **613**, 50–68 (2022)
47. M. Jagielski, G. Severi, N.P. Harger, A. Oprea, Subpopulation data poisoning attacks. In Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security. Association for Computing Machinery. New York. (2021). 3104-3122 https://dl.acm.org/doi/proceedings/10.1145/3460120
48. I. Rosenberg, A. Shabtai, Y. Elovici, L. Rokach, Adversarial machine learning attacks and defense methods in the cyber security domain. ACM Comput. Surv. **54**, (5) 1–36 (2021)
49. M. Goldblum, D. Tsipras, C. Xie, X. Chen, A. Schwarzschild, D. Song, A. Madry, B. Li, T. Goldstein, Dataset security for machine learning: Data poisoning, backdoor attacks, and defenses. IEEE Transactions on Pattern Analysis and Machine Intelligence. **45**,(2) 1563-1580 (2022)
50. M. Rigaki, S. Garcia, A survey of privacy attacks in machine learning. ACM Computing Surveys. **56**, (4) 1-34 (2023)
51. Z. Wang, J. Ma, X. Wang, J. Hu, Z. Qin, K. Ren, Threats to training: A survey of poisoning attacks and defenses on machine learning systems. ACM Comput. Surv. **55**, (7) 1–36 (2022)

52. N. Pitropakis, E. Panaousis, T. Giannetsos, E. Anastasiadis, G. Loukas, A taxonomy and survey of attacks against machine learning. Comput. Sci. Rev. **34**, 100199 (2019)
53. A. Shafee, T.A. Awaad, Privacy attacks against deep learning models and their countermeasures. J. Syst. Archit. **114**, 101940 (2021)
54. P. Bountakas, A. Zarras, A. Lekidis, C. Xenakis, Defense strategies for adversarial machine learning: A survey. Comput. Sci. Rev. **49**, 100573 (2023)
55. N. Martins, J.M. Cruz, T. Cruz, P. Henriques Abreu, Adversarial machine learning applied to intrusion and malware scenarios: A systematic review. IEEE Access **8**, 35403–35419 (2020)
56. G.R. Machado, E. Silva, R.R. Goldschmidt, Adversarial machine learning in image classification: A survey toward the defender's perspective. ACM Comput. Surv. **55**, (1) 1–38 (2021)
57. A. Alotaibi, M.A. Rassam, Adversarial machine learning attacks against intrusion detection systems: A survey on strategies and defense. Fut. Internet **15**,(2) 62 (2023)
58. A. Demontis, M. Melis, M. Pintor, M. Jagielski, B. Biggio, A. Oprea, C. Nita-Rotaru, F. Roli, On the intriguing connections of regularization, input gradients and transferability of evasion and poisoning attacks. arXiv preprint arXiv:1809.02861. (2018)
59. M. Jagielski, A. Oprea, B. Biggio, C. Liu, C. Nita-Rotaru, B. Li, Manipulating machine learning: Poisoning attacks and countermeasures for regression learning. In 2018 IEEE symposium on security and privacy (SP). IEEE. 19–35 (2018)
60. D. Gibert, J. Planes, Q. Le, G. Zizzo. Query-free evasion attacks against machine learning-based malware detectors with generative adversarial networks (2023)
61. H. Yan, X. Li, W. Zhang, R. Wang, H. Li, X. Zhao, F. Li, X. Lin, A wolf in sheep's clothing: Query-free evasion attacks against machine learning-based malware detectors with generative adversarial networks. In 2023 IEEE European Symposium on Security and Privacy Workshops (EuroS&PW) **21**, 415-426 (2023) IEEE. https://ieeexplore.ieee.org/xpl/conhome/10190553/proceeding
62. H. Bostani, V. Moonsamy, Evadedroid: A practical evasion attack on machine learning for blackbox android malware detection. Computers & Security. **139**, 103676–103693 (2024)
63. Y. Shi, Y.E. Sagduyu, in *MILCOM 2017 - 2017 IEEE Military Communications Conference (MILCOM)*. Evasion and causative attacks with adversarial deep learning. IEEE. 243–248 (2017). https://www.ieee.org/conferences/publishing/index.html
64. T. Titcombe, A.J. Hall, P. Papadopoulos, D. Romanini, Practical defences against model inversion attacks for split neural networks. arXiv preprint arXiv:2104.05743. (2021)
65. M. Khosravy, K. Nakamura, Y. Hirose, N. Nitta, N. Babaguchi, Model inversion attack: Analysis under gray-box scenario on deep learning based face recognition system. KSII Trans. Internet Inf. Syst. **15**, 1100–1119 (2021)
66. Q. Zhang, J. Ma, Y. Xiao, J. Lou, L. Xiong, in *2020 IEEE International Conference on Big Data (Big Data)*. Broadening differential privacy for deep learning against model inversion attacks. IEEE. 1061–1070 (2020). https://www.ieee.org/conferences/publishing/index.html
67. Z. He, T. Zhang, R.B. Lee, Model inversion attacks against collaborative inference. In Proceedings of the 35th Annual Computer Security Applications Conference publisher address: Association for Computing Machinery. New York. 148–162 (2019) https://dl.acm.org/doi/proceedings/10.1145/3359789
68. S. Basu, R. Izmailov, C. Mesterharm, Membership model inversion attacks for deep networks. arXiv preprint arXiv:1910.04257 (2019)
69. U. Aïvodji, S. Gambs, T. Ther, Gamin: An adversarial approach to black-box model inversion. arXiv preprint arXiv:1909.11835. (2019)
70. Z. Zhu, C. Wu, R. Fan, D. Lian, E. Chen, in *Proceedings of the ACM Web Conference 2023*. Membership inference attacks against sequential recommender systems (Association for Computing Machinery, 2023), pp. 1208–1219 https://www.ieee.org/conferences/publishing/index.html
71. J. Chen, W.H. Wang, X. Shi, Membership Inference Attacks Against Sequential Recommender Systems. In Proceedings of the ACM Web Conference publisher address: Association for Computing Machinery. New York. 1208–1219 (2023) https://dl.acm.org/doi/proceedings/10.1145/3543507
72. J. Chen, W.H. Wang, X. Shi, Differential privacy protection against membership inference attack on machine learning for genomic data. In BIOCOMPUTING 2021: Proceedings of the Pacific Symposium. World Scientific Publishing Company. 26–37 (2020) https://www.proceedings.com/58564.html
73. M. Zhang, Z. Ren, Z. Wang, P. Ren, Z. Chen, P. Hu, Y. Zhang, Membership inference attacks against recommender systems. In Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security. New York. (2021) https://dl.acm.org/doi/proceedings/10.1145/3460120
74. Y. Zou, Z. Zhang, M. Backes, Y. Zhang, Privacy analysis of deep learning in the wild: Membership inference attacks against transfer learning. arXiv preprint arXiv:2009.04872. (2020)
75. M. Khosravy, K. Nakamura, Y. Hirose, N. Nitta, N. Babaguchi, Model inversion attack by integration of deep generative models: Privacy-sensitive face generation from a face recognition system. IEEE Trans. Inf. Forensic Secur. 357–372 (2022)
76. S. Venkatesan, H. Sikka, R. Izmailov, R. Chadha, A. Oprea, M.J. de Lucia, in *MILCOM 2021 - 2021 IEEE Military Communications Conference (MILCOM)*. Poisoning attacks and data sanitization mitigations for machine learning models in network intrusion detection systems (2021), pp. 874–879
77. S. Ho, A. Reddy, S. Venkatesan, R. Izmailov, R. Chadha, A. Oprea, in *MILCOM 2022 - 2022 IEEE Military Communications Conference (MILCOM)*. Data sanitization approach to mitigate clean-label attacks against malware detection systems. IEEE. 993–998 (2022). https://www.ieee.org/conferences/publishing/index.html
78. A. Paudice, L. Mu noz-González, E.C. Lupu, in *ECML PKDD 2018 Workshops: Nemesis 2018, UrbReas 2018, SoGood 2018, IWAISe 2018, and Green Data Mining 2018, Dublin, Ireland, September 10-14, 2018, Proceedings 18*. Label sanitization against label flipping poisoning attacks (Springer, 2018), pp. 5–15 https://link.springer.com/book/10.1007/978-3-030-13453-2
79. P.W. Koh, J. Steinhardt, P. Liang, Stronger data poisoning attacks break data sanitization defenses. Machine Learning. **111**, 1–47 (2022)
80. P.P. Chan, Z.M. He, H. Li, C.C. Hsu, Data sanitization against adversarial label contamination based on data complexity. Int. J. Mach. Learn. Cybern. **9**, 1039–1052 (2018)
81. T.Y. Liu, Y. Yang, B. Mirzasoleiman, Friendly noise against adversarial noise: a powerful defense against data poisoning attack. Advances in Neural Information Processing Systems. **35**, 11947–11959 (2022)
82. Z. You, J. Ye, K. Li, Z. Xu, P. Wang, in *2019 IEEE International Conference on Image Processing (ICIP)*. Adversarial noise layer: Regularize neural network by adding noise. IEEE. 909–913 (2019). https://www.ieee.org/conferences/publishing/index.html
83. G. Tao, Y. Liu, G. Shen, Q. Xu, S. An, Z. Zhang, X. Zhang, in *2022 IEEE Symposium on Security and Privacy (SP)*. Model orthogonalization: Class distance hardening in neural networks for better security. IEEE. 1372–1389 (2022). https://www.ieee.org/conferences/publishing/index.html
84. G. Apruzzese, M. Andreolini, M. Colajanni, M. Marchetti, Hardening random forest cyber detectors against adversarial attacks. IEEE Trans. Emerg. Top. Comput. Intell. **4**(4), 427–439 (2020)
85. M. Pawlicki, M. Choraś, R. Kozik, Defending network intrusion detection systems against adversarial evasion attacks. Futur. Gener. Comput. Syst. **110**, 148–154 (2020)
86. U. Ahmed, J.C.W. Lin, G. Srivastava, Mitigating adversarial evasion attacks by deep active learning for medical image classification. Multimed. Tools Appl. **81**(29), 41899–41910 (2022)
87. H. Rafiq, N. Aslam, U. Ahmed, J.C.W. Lin, Mitigating malicious adversaries evasion attacks in industrial internet of things. IEEE Trans. Ind. Inform. **19**(1), 960–968 (2023)
88. J. Lin, L.L. Njilla, K. Xiong, Secure machine learning against adversarial samples at test time. EURASIP J. Inf. Secur. **2022**(1), 1 (2022)
89. G. Apruzzese, M. Andreolini, M. Marchetti, A. Venturi, M. Colajanni, Deep reinforcement adversarial learning against botnet evasion attacks. IEEE Trans. Netw. Serv. Manag. **17**(4), 1975–1987 (2020)
90. E. Anthi, L. Williams, A. Javed, P. Burnap, Hardening machine learning denial of service (dos) defences against adversarial attacks in iot smart home networks. Comput. Secur. **108**, 102352 (2021)
91. X. Cao, N.Z. Gong, in *Proceedings of the 33rd Annual Computer Security Applications Conference*. Mitigating evasion attacks to deep neural

networks via region-based classification. JMLR. 278–287 (2017). https://jmlr.org/
92. H. Phan, M.T. Thai, H. Hu, R. Jin, T. Sun, D. Dou, in *Proceedings of the 37th International Conference on Machine Learning*. Scalable differential privacy with certified robustness in adversarial learning, vol. 119 (PMLR, 2020), pp. 7683–7694 JMLR.org
93. M. Strobel, R. Shokri, Data privacy and trustworthy machine learning. IEEE Secur. Priv. **20**(5), 44–49 (2022)
94. K. Pan, M. Gong, Y. Gao, Privacy-enhanced generative adversarial network with adaptive noise allocation. Knowl. Based Syst. **272**, 110576 (2023)
95. J. Jin, E. McMurtry, B.I.P. Rubinstein, O. Ohrimenko, in *2022 IEEE Symposium on Security and Privacy (SP)*. Are we there yet? timing and floating-point attacks on differential privacy systems. IEEE. 473–488 (2022). https://www.ieee.org/conferences/publishing/index.html
96. Z. Chen, J. Wu, A. Fu, M. Su, R.H. Deng, Mp-clf: An effective model-preserving collaborative deep learning framework for mitigating data leakage under the gan. Knowl. Based Syst. **270**, 110527 (2023)
97. Z. Chen, K. Pattabiraman, Overconfidence is a dangerous thing: Mitigating membership inference attacks by enforcing less confident prediction. arXiv preprint arXiv:2307.01610 (2023)
98. Z. Yang, L. Wang, D. Yang, J. Wan, Z. Zhao, E.C. Chang, F. Zhang, K. Ren, in *Proceedings of the AAAI Conference on Artificial Intelligence*. Purifier: Defending data inference attacks via transforming confidence scores. The AAAI Press. **37**, 10871–10879 (2023). https://aaai.org/aaai-publications/
99. Z. Zhang, Q. Liu, Z. Huang, H. Wang, C.K. Lee, E. Chen, Model inversion attacks against graph neural networks. IEEE Trans. Knowl. Data Eng. **35**(9), 8729–8741 (2023)
100. T. Zhu, D. Ye, S. Zhou, B. Liu, W. Zhou, Label-only model inversion attacks: Attack with the least information. IEEE Trans. Inf. Forensic Secur. **18**, 991–1005 (2023)
101. Y. Liu, Z. Zhao, M. Backes, Y. Zhang, Membership inference attacks by exploiting loss trajectory. In Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security. Association for Computing Machinery. New York. 2085–2098 (2022) https://dl.acm.org/doi/proceedings/10.1145/3548606
102. L. Liu, Y. Wang, G. Liu, K. Peng, C. Wang, Membership inference attacks against machine learning models via prediction sensitivity. IEEE Trans. Dependable Secure Comput. **20**(3), 2341–2347 (2023)
103. N. Carlini, S. Chien, M. Nasr, S. Song, A. Terzis, F. Tramèr, in *2022 IEEE Symposium on Security and Privacy (SP)*. Membership inference attacks from first principles. IEEE. 1897–1914 (2022). https://www.ieee.org/conferences/publishing/index.html
104. R.S. Siva Kumar, M. Nyström, J. Lambert, A. Marshall, M. Goertzel, A. Comissoneru, M. Swann, S. Xia, in *2020 IEEE Security and Privacy Workshops (SPW)*. Adversarial machine learning-industry perspectives. IEEE. 69–75 (2020). https://www.ieee.org/conferences/publishing/index.html
105. M. Schreyer, T. Sattarov, B. Reimer, D. Borth, Adversarial learning of deepfakes in accounting. (2019). CoRR abs/1910.03810
106. I. Fursov, M. Morozov, N. Kaploukhaya, E. Kovtun, R. Rivera-Castro, G. Gusev, D. Babaev, I. Kireev, A. Zaytsev, E. Burnaev, *Adversarial attacks on deep models for financial transaction records.* In Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining. Association for Computing Machinery, New York, NY, United States. 2868–2878 (2021) https://dl.acm.org/doi/proceedings/10.1145/3447548
107. A. Rahman, M.S. Hossain, N.A. Alrajeh, F. Alsolami, Adversarial examples-security threats to covid-19 deep learning systems in medical iot devices. IEEE Internet Things J. **8**(12), 9603–9610 (2021)
108. X. Han, Y. Hu, L. Foschini, L. Chinitz, L. Jankelson, R. Ranganath, Deep learning models for electrocardiograms are susceptible to adversarial attack. Nat. Med. **26**(3), 360–363 (2020)
109. X. Ma, Y. Niu, L. Gu, Y. Wang, Y. Zhao, J. Bailey, F. Lu, Understanding adversarial attacks on deep learning based medical image analysis systems. Pattern Recogn. **110**, 107332 (2021)
110. H. Kim, D.C. Jung, B.W. Choi, Exploiting the vulnerability of deep learning-based artificial intelligence models in medical imaging: adversarial attacks. J. Korean Soc. Radiol. **80**(2), 259–273 (2019)
111. G. Apruzzese, M. Colajanni, L. Ferretti, M. Marchetti, in *2019 11th international conference on cyber conflict (CyCon)*. Addressing adversarial attacks against security systems based on machine learning, vol. 900 (IEEE, 2019), pp. 1–18 https://ieeexplore.ieee.org/xpl/conhome/8751947/proceeding
112. A. Piplai, S.S.L. Chukkapalli, A. Joshi, in *2020 IEEE 6th Intl Conference on Big Data Security on Cloud (BigDataSecurity). IEEE Intl Conference on High Performance and Smart Computing,(HPSC) and IEEE Intl Conference on Intelligent Data and Security (IDS)*, Nattack! adversarial attacks to bypass a gan based classifier trained to detect network intrusion (IEEE, 2020), pp. 49–54 https://www.ieee.org/conferences/publishing/index.html
113. A. Kuppa, N.A. Le-Khac, in *2020 International Joint Conference on neural networks (IJCNN)*. Black box attacks on explainable artificial intelligence (xai) methods in cyber security (IEEE, 2020), pp. 1–8 https://www.ieee.org/conferences/publishing/index.html
114. E. Raff, M. Benaroch, A.L. Farris, You Don't Need Robust Machine Learning to Manage Adversarial Attack Risks. arXiv preprint arXiv:2306.09951 (2023)

Publisher's Note