Review Article

Ghadeer Ghazi Shayea, Mohd Hazli Mohammed Zabil, Mustafa Abdulfattah Habeeb, Yahya Layth Khaleel, and A. S. Albahri*

Strategies for protection against adversarial attacks in AI models: An in-depth review

https://doi.org/10.1515/jisys-2024-0277
received May 27, 2024; accepted August 19, 2024

Abstract: The enhanced use of artificial intelligence (AI) in organizations has changed and revolutionized the approaches to solving problems, processing information, and decision making. While the algorithms turned out to be highly effective, AI systems faced adversarial attacks, which can be described as slight alterations of inputs that would fool an AI algorithm. These attacks remain major challenges to the dependability and protection of AI systems and thus the need to develop stable and flexible protection strategies and procedures. The aim of this article is to discuss the existing trends in adversarial attack techniques and protection mechanisms. To this end, papers, exact match, and systematically applied operative inclusion/exclusion criteria pertinent to protection strategies against adversarial attacks in multiple databases were incorporated and used. Specifically, 1988 papers were retrieved from Web of Science, IEEE Explore, and Science Direct, which were published between January 1, 2021, and July 1, 2024, where we used 51 of the identified journal articles for the quantitative synthesis in the final stage. Thus, the protection taxonomy, which resulted from our analysis, discusses the motivation, and best practices in relation to the threats in question. The taxonomy also describes challenges and suggests other ideas on how to improve the robustness of adversarial attack systems. Not only this study is a response to gaps in the literature but it also presents the reader with a map for further studies. It is necessary to draw attention to the fact that an objective criterion must be introduced to measure the degree of defense, collaboration with researchers of other fields, and the necessity to consider the ethical implications of the created defense mechanisms. Our results shall assist industry practitioners, researchers, and policymakers in designing an optimal AI security that can protect AI systems against dynamic adversary strategies. This review provides a reference of entry to the topic of AI security and the challenges that may be encountered together with the measures that can be taken to forward the studies.

Keywords: artificial intelligence, adversarial attack, security, algorithms, strategies, machine learning, deep learning

*** Corresponding author: A. S. Albahri,** Department of Computer Technology Engineering, Technical College, Imam Ja'afar Al-Sadiq University, Baghdad, 10001, Iraq; Electronic Computer Center, University of Information Technology and Communications (UoITC), Baghdad 10013, Iraq, e-mail: ahmed.albahri@ijsu.edu.iq
Ghadeer Ghazi Shayea: College of Information Technology, Universiti Tenaga Nasional (UNITEN), 43000, Kajang, Selangor, Malaysia; Technical College, Imam Ja'afar Al-Sadiq University, Baghdad, 10001, Iraq, e-mail: ghadeer.ghazi@ijsu.edu.iq
Mohd Hazli Mohammed Zabil: College of Computing & Informatics, Universiti Tenaga Nasional, Jalan IkRAM – UNITEN, 43000 Kajang, Malaysia, e-mail: hazli@uniten.edu.my
Mustafa Abdulfattah Habeeb: Department of Computer Science, College of Computer Science and Mathematics, Tikrit University, Salah Al Deen, 34001, Iraq, e-mail: mustafa@tu.edu.iq
Yahya Layth Khaleel: Department of Computer Science, College of Computer Science and Mathematics, Tikrit University, Salah Al Deen, 34001, Iraq, e-mail: yahya@ztu.edu.iq

1 Introduction

The ultimate goal of artificial intelligence (AI) is achieved through data and algorithm integration, with the results that the AI system produces being even more effective than the ones made by humans. Facial recognition, the voice of virtual assistance speech recognition, and self-driving cars are current examples of AI technology that we are accustomed to [1,2]. This type of human-less intelligence is called AI. The ability to recognize one's surroundings and decide what to do independently or any device or machine that can replicate the human brain's mental processes is referred to as AI. Together, machine learning (ML) and deep learning (DL) create a family of AI methods that can be called subsets of AI [3–5]. Therefore, ML is a part of the AI branch that focuses on enabling computations to self-learn, meaning that no direct programming is needed. The use of ML entails designing algorithms that can learn presumably from data and then predict the data [6,7]. As AI has dramatically progressed and spread in the past decade, the number of DL applications in industrial Internet-of-thing (IIoT) ecosystems will likely sharply increase over the current decade [8]. However, in recent years, the pervasive integration of ML and DL systems across various domains has revolutionized complex problem solving [9], decision making, and data analysis. As ML and DL algorithms continue to demonstrate remarkable capabilities, they have also become susceptible to adversarial attacks, a burgeoning concern that has captured the attention of researchers, practitioners, and industry experts alike [10,11]. Adversarial attacks are a serious risk to the resilience and dependability of ML and DL models, necessitating a profound exploration of protection strategies and methods [12–14].

This systematic review embarks on a comprehensive journey to dissect and analyze the multifaceted landscape of protection mechanisms deployed to safeguard ML and DL systems against adversarial attacks. Adversarial attacks, often subtle manipulations of input data with the intent of misleading or subverting ML models, have revealed vulnerabilities even in the most sophisticated algorithms. Such attacks have manifested in various domains, including computer vision [15], natural language processing (NLP) [16], and reinforcement learning [17], underscoring the urgency of robust defenses. Our primary objectives in this expansive study are to identify, categorize, and critically evaluate the diverse range of defense strategies and methods proposed in the existing body of literature. By synthesizing insights from a multitude of research papers, articles, and conference proceedings, the goal of this systematic review is to present a comprehensive and organized overview of the most recent state-of-the-art protection in ML and DL against adversarial attacks. The journey begins with an exploration of the fundamental principles underlying adversarial attacks, unraveling the intricate interplay between attackers and defenders in the ML and DL landscapes. We navigate through the taxonomy of adversarial attacks, ranging from evasion attacks [18] that manipulate input data to poisoning attacks [19] that compromise the integrity of training data. Understanding an adversary's arsenal is pivotal for crafting effective protection mechanisms that can be resilient in the face of evolving threats.

Moving forward, the review systematically categorizes protection strategies on the basis of their methodologies, encompassing approaches such as adversarial training (AT), input preprocessing, robust optimization, and model ensembling. Each category is dissected to unravel the underlying mechanisms, strengths, and limitations, providing a nuanced perspective on the efficacy of various protection strategies. Furthermore, we examine the intersection of these protection strategies with explainable, scalable, and real-world applicability, acknowledging the multifaceted challenges that emerge when deploying protection in practical settings. Throughout this exploration, we emphasize the dynamic nature of the adversarial landscape, emphasizing the need for adaptive protection strategies that evolve in tandem with emerging attack methodologies. The insights garnered from this review not only contribute to a deeper understanding of protection mechanisms but also clarify the path for future research directions and the development of robust, secure, and resilient ML and DL systems. In conclusion, this systematic review serves as a comprehensive guide for researchers, practitioners, and policymakers engaged in the realm of ML and DL security. By synthesizing knowledge from diverse sources, we focus on fostering an understanding of the current state-of-the-art protection against adversarial attacks, catalyzing advancements in developing trustworthy and secure ML systems.

The primary axis, which is usually considered in the majority of studies concerned with adversarial attacks, is the protection aspect. In this systematic literature review, we aim to shed light upon all the dimensions of protection approaches and further evaluate their effectiveness in terms of protecting the ML/DL models. Thus, the mission of this work is to make a small but significant step toward enhancing the protection of AI systems and

to support the constant development of effective and robust precautions and protections in the fields in which AI is used (Figure 1).

Figure 1: The divisions of protection systems in adversarial attacks. Source: Created by the authors.

Therefore, according to our opinion and analysis, protection systems include the following:

- Defense methods: Advanced preventive actions, such as AT and robust optimization, are intended to identify possible attacks and safeguard against them.
- Training on attacks: Introduction of "adversarial" patterns into the training procedure to increase model robustness.
- Intrusion detection: Different strategies for distinguishing an attack through an examination of data and the use of smart algorithms.

The research objectives aligned with these motivations are articulated through the following questions:

Q1: What defines an appropriate taxonomy for incorporating protection strategies and methods in AI models against adversarial attacks?

Q2: How do studies address the integration of protection strategies concerning motivations, challenges, recommendations, and limitations?

Q3: What are the notable gaps in the current research on methods of protection *versus* adversarial attacks in AI methods?

The key contributions of this study are as follows:

(1) Pioneering a thorough exploration of adversarial attacks in the innovation and optimization of protection methods, generation of adversarial attacks, protection strategies, protection robustness, and applications in malware, intrusion, and anomaly detection.
(2) This article provides an extensive literature review that deeply examines prevailing research trends, challenges, motivations, limitations, and recommendations within the domain of protection strategies and methods against adversarial attacks in ML and DL.
(3) Identifying research gaps, proposing future directions and outlining a roadmap to enhancing protection strategies across diverse scenarios of adversarial attacks.

The article's structure unfolds as follows: Section 2 presents an analysis of adversarial attacks. Section 3 presents the methodology used in the study, and Section 4 outlines the bibliometric analysis of the literature as well as an in-depth analysis of the papers. In addition, the taxonomy results related to protection strategies are explored, and Section 5 discusses motivations, challenges, and recommendations. Section 6 provides a critical analysis, pinpointing gaps in current research and proposing distinct research avenues. Finally, Section 7 summarizes and concludes this study.

2 Adversarial attacks: An overview

The use of ML has been on the rise in the recent past, cutting across almost all disciplines. DL, in particular, has been actively used and continues to act as the basis for applications and services in areas such as computer vision, language processing, translation, and security, whose functions are sometimes better than those at the human level [20]. Nevertheless, the suggested systems are still vulnerable to adversarial attacks. By altering inputs with the goal of either evading detection or changing the model's classification, adversarial attacks are a significant threat in many applications, such as self-driving cars and health care [21]. Indeed, security threats and their countermeasures from the aspect of adversarial attacks have attracted much attention in recent years. This is because these attacks can compromise the credibility and reliability of ML solutions. Therefore, the field has been oriented toward examining several aspects of these problems. Despite this, there remains a notable gap in the literature, which lacks a distinct classification that can generally describe the adversarial process. Such frameworks are necessary to have a systematic approach to studying adversarial attacks and, therefore, enable the scientific establishment of defenses against them. Therefore, it remains a constant necessity to develop and share reliable and easily discussed graphic classification and division methods on

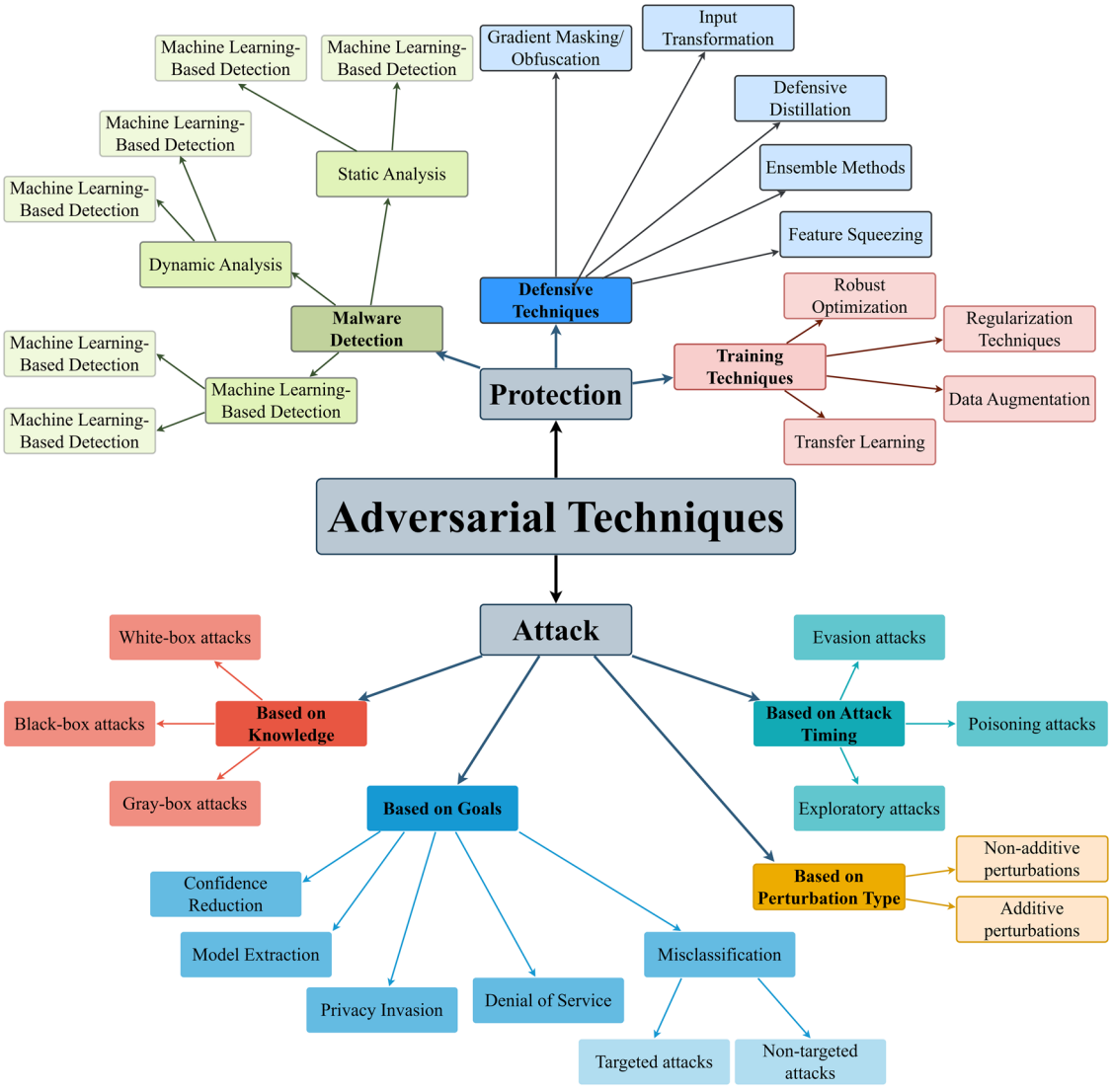

Figure 2: Adversarial techniques. Source: Created by the authors.

the basis of the forms of adversarial attacks and protection. Adversarial attacks can be classified into many forms, as in previous studies [22,23]. Thus, according to numerous previous studies, we have proposed a general diagram that explains the concept of adversarial techniques in our opinion. This diagram is useful in the literature review of other related research in the future and might help people better understand what adversarial is and categorize it accordingly to further develop the theoretical background of this field. The diagram is presented in Figure 2.

As we note in Figure 2, adversarial techniques can be classified into two basic categories: protection and attack, which are as follows:

Adversarial attacks

1. Based on Knowledge
   - White-box attacks: The attacker knows the architecture and parameters of the model as well as the data used during training [24].
   - Black-box attacks: When the model of the attacker is fully unknown, only the input and output [25] are known.
   - Gray-box attacks: The attacker can design attacks based on knowledge of the model's architecture but without knowledge of the specific parameters [26].
2. Based on perturbation type
   - Additive perturbations: Small, often imperceptible changes are added to the input data to deceive the model [27].
   - Nonadditive perturbations: It originated from modifying the input data by adding small, sometimes almost negligible changes to deceive the model [28].
3. Based on attack timing
   - Evasion attacks: There is a change in inputs at the time of inference to avoid detection/classification [29,30].
   - Poisoning attacks: The attacker intervenes with the training data during the training phase to affect the model [31,32].
   - Exploratory attacks: The attacker checks the model for weaknesses or any unusual response, which helps the attacker understand how it works [33].
4. Based on goals
   - Misclassification
     – Targeted attacks: The attacker wants the input to be categorized into a specific, wrong class that is completely different from its original class [34,35].
     – Nontargeted attacks: The attacker wants to introduce any desired misclassification with no need to direct at a specific class [35].
   - Confidence reduction: The attacker focuses on confusing the model into being less sure of its classifications even if the classification made is correct [36].
   - Model extraction: The attacker tries to obtain the model parameters or architecture by querying [37].
   - Privacy invasion: The attacker attempts to learn the specifics concerning the training data that have informed the creation of the model [38].
   - Denial of service: The attacker aims to reduce the effectiveness of the model or complete failure of the given model [39].

Adversarial protection

1. Defensive techniques
   - Gradient masking/obfuscation: This prevents the attacker from computing good-quality gradients by either changing the model or the loss function used [40].
   - Input transformation: Cleans the input data to eliminate or reduce the effect of an adversarial perturbation on the given data [41].
   - Defensive distillation: The model is trained at a high temperature, and the output is suitably softened to be used to train the last model; thus, generating adversarial examples becomes difficult for an attacker [42].

- Ensemble methods: Combines multiple models to increase robustness to decrease their vulnerability because attacking an ensemble is not as easy as attacking a single model [43,44].
- Feature squeezing: This simplifies the input features (for instance, by lowering the number of color bits in images) to restrict the ability of the attacker to create efficient perturbations [45].

2. Training techniques
   - Robust optimization: This entails optimizing the model parameters with the inherent trait of being largely acceptable for adversarial perturbations [46].
   - Regularization techniques: Regularization is used for models to decrease the problem of overfitting, such as L2 regularization or dropout [47].
   - Data augmentation: Data diversification can be used to increase the robustness of the model by adding new, variable examples into the training dataset [48].
   - Transfer learning: Training on models that have prior knowledge that has been trained on large, diverse datasets that in itself may contain more robustness [49].
3. Malware detection
   - Static analysis: Analyze the code and static structures of the files to look for malware [50].
     – Signature-based detection: Scans the found files with the bases containing the signatures of viruses and then reports the files [51].
     – Heuristic analysis: Utilizing the heuristic rules that let us look for the previously identified suspicious patterns in code [52].
   - Dynamic analysis: Oversees the files of the system when in the process of execution with the purpose of detecting malicious activities [50].
     – Behavioral analysis: Records various activities that take place at one or more phases of a program run, such as file update activities and Internet connections [53].
     – Anomaly detection: Identifies deviations from normal behavior that may indicate malicious activity [54].
   - ML detection: Determining the malware by identifying the genre of the malware on the basis of patterns and features from the data [55].
     – Feature extraction: Extract relevant features from the data, such as API calls, file attributes, and network traffic [56].
     – Classification models: Models such as decision trees [57], random forests [58], and neural networks [59] are used to classify files as “safe” or “malware.”

When these classifications are well understood, one can easily embrace the fact that adversarial attacks are quite diverse and complicated and be in a better position to design appropriate protection measures for ML systems against such threats. In this context, the focus of this work will be on protection against adversarial attacks on the basis of a discussion of relevant research contributions and possible advances. This study will seek to illustrate ways through which different measures have been utilized in protection against adversarial threats and how further development could be made in relation to the topic at hand with respect to ML. Thus, our systematic review aims to help advance research in this essential field and inspire innovative approaches.

3 Methodology

This study adhered to the methodology used in prior research [60,61] and included a systematic literature review using the preferred reporting items for systematic reviews and meta-analysis (PRISMA) statement. The analysis section followed the recommended reporting guidelines for systematic reviews and meta-analyses [60,61]. Various bibliographic citation databases covering a range of scientific and social science journals across different disciplines were used for the research. The quest for relevant papers included searching four widely recognized and reliable digital databases: Science Direct (SD), Scopus, IEEE Xplore (IEEE), and Web of Science (WoS) [62–64]. These databases are crucial for researchers, offering extensive coverage of scientific and technological research and supplying valuable insights for further analysis and investigation.

3.1 Search strategy

The search process involved comprehensively exploring the four selected databases to gather academic publications published in English. The search scope encompassed all scientific publications from 2021 to 1 July 2024. To conduct the search, a Boolean query was implemented, utilizing the "AND" operator to connect the keywords "adversarial attack," "machine learning," and "deep learning" (refer to Figure 5 for the detailed query). The selection of these keywords was based on recommendations provided by experts in AI, ML, DL, and decision making. This approach aimed to ensure a comprehensive and focused search strategy for identifying relevant literature.

3.2 Inclusion and exclusion criteria

The inclusion or selection of papers was based on the following criteria:

- The papers had to be written in English and published in reputable journals or conference proceedings.
- The papers should include adversarial attacks utilizing AI models (ML or DL).
- The selected papers were required to address protection within the realm of adversarial attacks, as mentioned earlier.

The following exclusion criteria were applied:

- Papers discussing adversarial attacks in areas unrelated to AI were not considered, and vice versa.
- Studies focusing on adversarial attacks in ML and DL but lacking relevance to protection were excluded, and vice versa.

3.3 Study selection

This method comprises a series of structured steps, starting with the identification and elimination of papers in duplicate. The titles and abstracts of the selected articles were carefully evaluated via Mendeley software. This initial screening resulted in the exclusion of numerous unrelated works, ensuring a focus on relevant literature. In instances of differences or inconsistencies among authors' assessments, the corresponding author played a crucial role in resolution and consensus. The subsequent step involved a thorough examination of the full texts of the articles, which were meticulously evaluated against the previously defined inclusion criteria in Section 3.2. This step aimed to refine the selection process by excluding articles that did not meet the predetermined criteria. The process and its outcomes are depicted in Figure 3, which provides an overview of the steps in filtering and selecting the final set of articles for analysis.

In this research, they focused on identifying and selecting those articles that met a set of specified criteria. First, a comprehensive search revealed 1988 entries comprising the articles from the SD, totaling 1,111; furthermore, 545 in Scopus, 130 in IEEE, and 202 in WoS. For the elimination of redundancy, 287 papers were found and removed, leaving no remaining number of papers at this number (1,701). Therefore, detailed scrutiny of titles and abstracts revealed that 1,073 articles were excluded because they did not comply with the predefined yardsticks. A comprehensive analysis was then performed for the subsequent 628 contributions. A total of 577 studies were excluded from the research because they failed to meet other inclusion criteria; 51 of these studies were excluded because they were determined to be relevant to basic inode requirements. In the end, 51 of these studies were included in the final collection of articles.

4 Finding analysis

By making a systematic attempt at categorization and analysis, this research aims to provide insightful knowledge regarding protection strategies and methods related to how they prevail during adversarial attacks. The

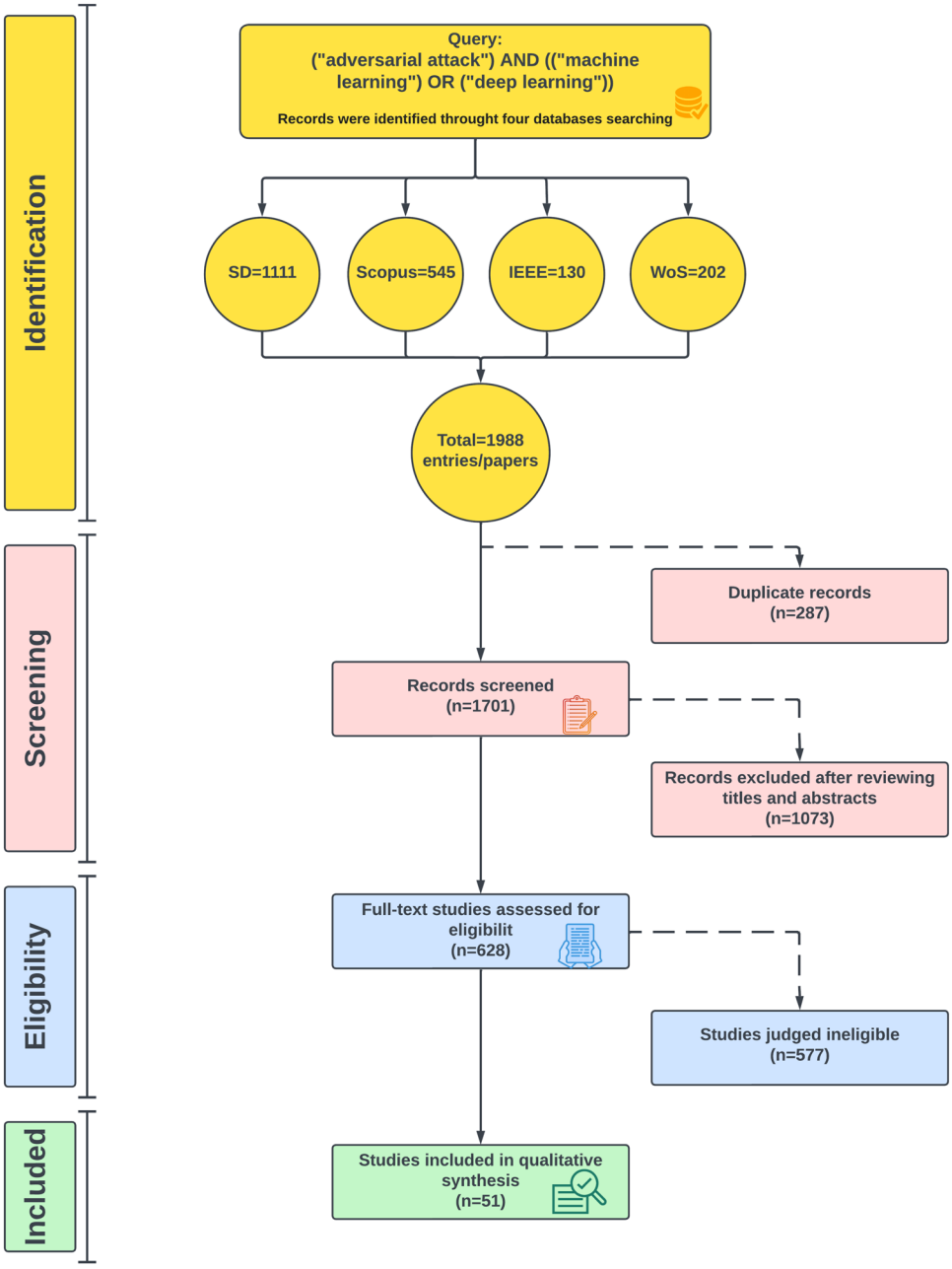

Figure 3: SLR protocol of protection strategies against adversarial attacks in AI models. Source: Created by the authors.

conclusive set of findings in the articles is explained in Section 4.1, where a comprehensive analysis and segregation occur, dividing them into separate categories on the basis of their specific objectives as well as contributions that they make to this perspective piece. This section provides a summary of the key findings and insights gained from the selected articles, which helps to better comprehend protection strategies and methods in terms of adversarial attacks.

4.1 Bibliometric analysis

The large number of papers has made us struggle to understand classic writings in our previous studies. Currently, owing to the availability of thousands of guides and articles, a large amount of information is available, which is rather difficult to follow. Some scholars support the PRISMA framework, suggesting the replacement of previous ones with an elaboration of problems, identification of research gaps, and theory development. In addition, even though systematic reviews provide much evidence, they might also contribute to the emergence of research paradigms and literature products, but they still have reliability and objectivity issues. This occurs from the authors' opinions, which they depend on to rephrase earlier knowledge. To add transparency to the treatment of summarizing past study results, many study projects have promoted holistic science mapping analysis via the RStudio package. The application of the bibliometric approach produces undeniable findings that reveal all the unfolding scenarios drawn from literary and scientific materials with high clarity and trustworthiness. In addition, the suggested tools are simple to use and free of charge, and no special competencies are needed. Consequently, this study makes use of one of the more elaborate bibliometric approaches that are illustrated in the following sections.

4.1.1 Most relevant sources

Figure 4 summarizes the most popular journals as sources of publications appending to the number of citations for each journal. The spot demonstrates the most influential and most often cited journals that were used for papers.

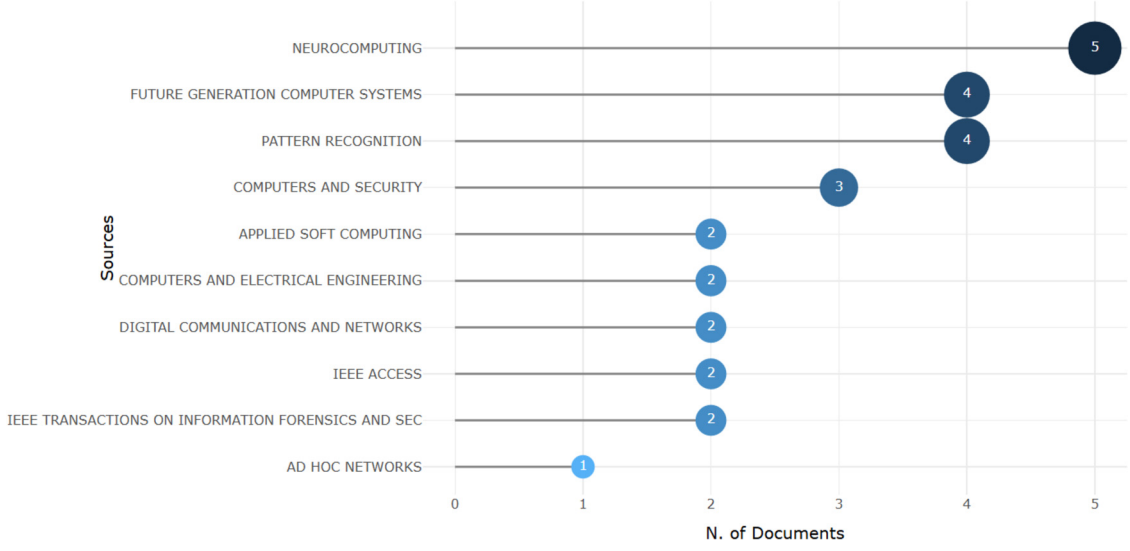

Figure 4: Most relevant sources. Source: Created by the authors.

4.1.2 Words' frequency over time

The top words found in the titles of the research papers that have been used in this research or even in their abstracts are shown in Figure 5.

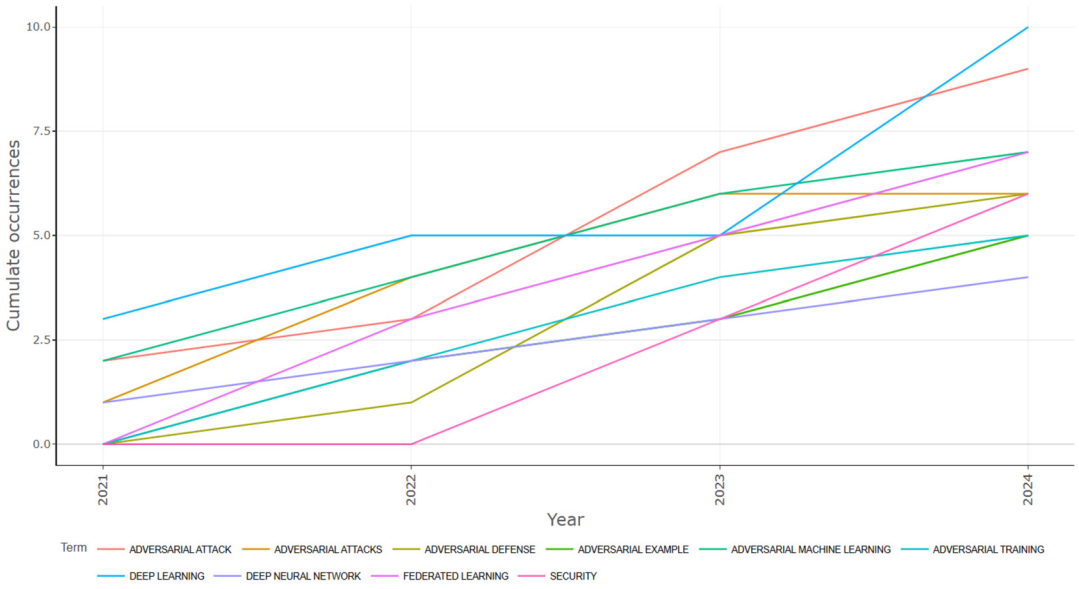

Figure 5: Words' frequency over time. Source: Created by the authors.

4.1.3 Word cloud

The word cloud diagram shown in Figure 6 comprises the most commonly occurring words from the documents discussing protection strategies and methods of ML and DL, which work to fight off adversarial attacks [65]. This finding was expressed in the term "adversarial attack," which appears most often in this context.

Figure 6: Word cloud. Source: Created by the authors.

4.1.4 Tree map

The depiction of hierarchical data in a conventional way is carried out via a directed tree structure. On the other hand, many trees cannot be displayed in small environments. In this fashion, the tree-map algorithm was conceived to render more than thousands of nodes in a more efficient manner [66]. Figure 7 shows the conceptual layout of the study.

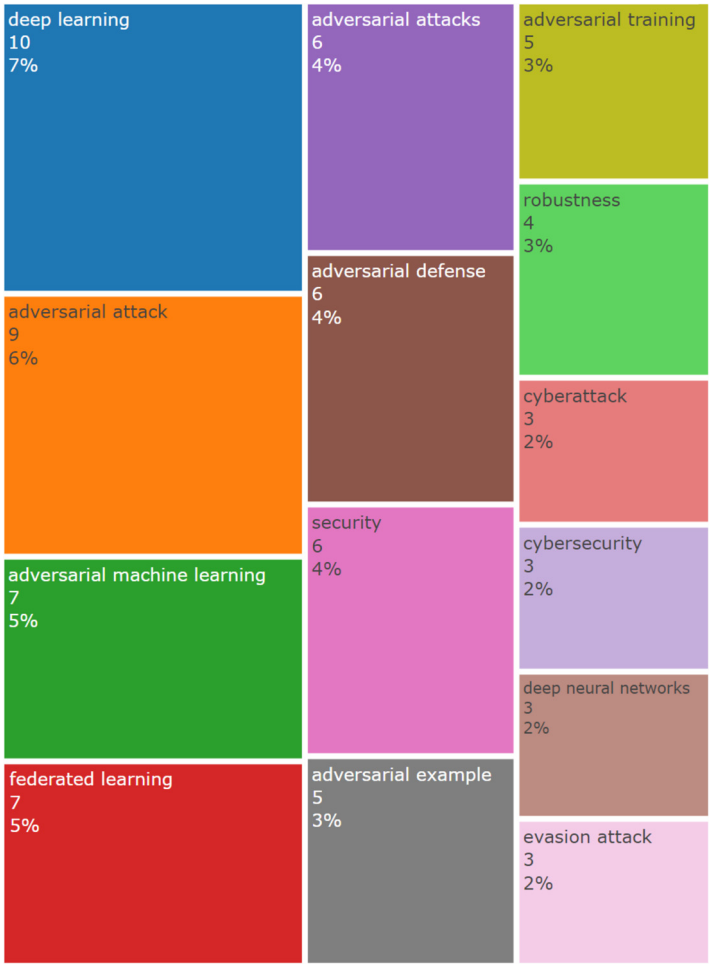

Figure 7: Tree map. Source: Created by the authors.

4.1.5 Cooccurrence network

The fundamental research method within bibliometric analysis is co-occurrence networks which experts use to investigate phenomena. The wide network of important concepts emerges from the analysis through previous linked terms to deliver a conceptual framework to policy experts and professionals about the investigated field [62]. Figure 8 shows the co-occurrence network created from the articles of the study.

4.1.6 Thematic map

The density and centrality indices served as foundation to create a thematic map which divided into four topological regions (Figure 9). The analysis determined this conclusion through surveys of the abstracts and titles from all analyzed references along with supplementing relevant keywords.

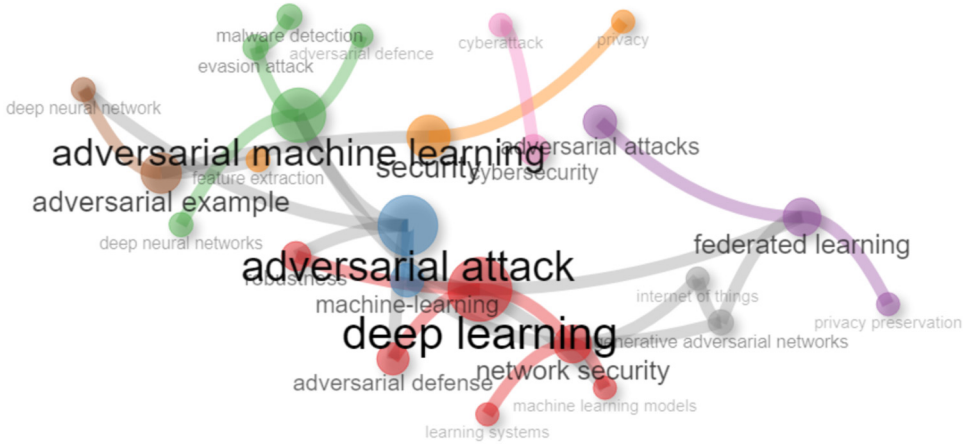

Figure 8: Co-occurrence network map. Source: Created by the authors.

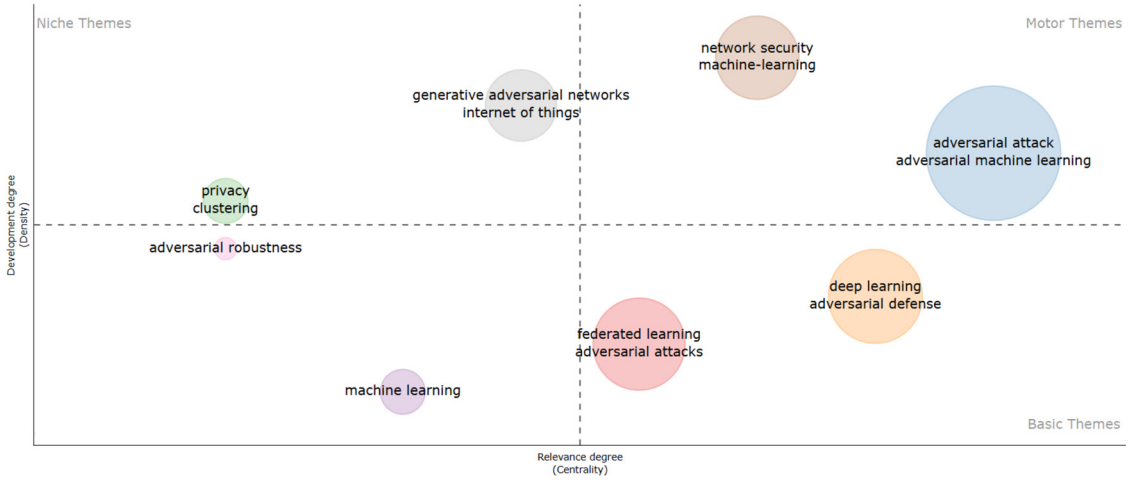

Figure 9: Thematic map. Source: Created by the authors.

4.1.7 Factorial analysis

Factorial analysis provides the similarity standardization capability for bibliographic coupling and cocitation and co-occurrence measures. Researchers operate this method to visualize discipline conceptual frameworks through frequency evaluation of bibliographic clusters [67] (see Figure 10).

4.2 Protection against adversarial attacks in AI models: Taxonomy

After the categorization of the 51 chosen articles, four groups were formulated to perform a systematic analysis on the basis of objective evidence and sourced from studies that fulfilled predetermined criteria. These study findings were primarily categorized into various subcategories to improve their organization and clarity during the presentation (Figure 11). These subdivisions allow the in-depth investigation of individual elements and the use of protection within adversarial attacks, thus enabling a holistic discussion of advancements and queasiness in this field. Furthermore, in the chosen articles, subcategories delve into other protection strategies and methods with respect to adversarial attacks that focus on what type of dataset has been used –

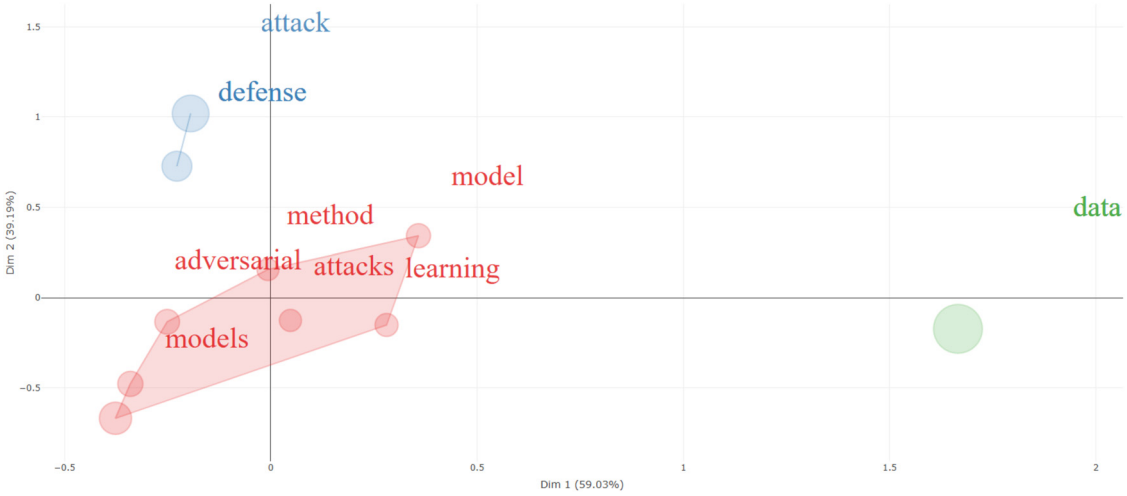

Figure 10: Factor analysis – three clusters. Source: Created by the authors.

Figure 11: Taxonomy of protection strategies in AI models against adversarial attacks. Source: Created by the authors.

whether text only – the image or a mix-up between both as far as this article is concerned. The established categories encompass n = 51 contributions, as outlined below:

(1) Innovation and optimization of defense methods: including 27 of 51 contributions.
(2) Malware, intrusion, and anomaly detection: 7 of 51 contributions.
(3) Generate adversarial attack and defense strategy: 7 of 51 contributions.
(4) Defense robustness: 10 of 51 contributions.

4.2.1 Innovation and optimization of defense methods

In the field of innovation and optimization of defense methods, a notable subset comprises 27 articles out of the 51 chosen, which explicitly delve into the implementation of defensive strategies and methodologies against adversarial attacks. These selected articles comprehensively investigate the approaches employed in various papers, analyzing how they leverage datasets, whether in the form of textual data, image data, or the integration of both modalities in their research endeavors.

Two studies delve into the innovation and optimization of defense methods in the text. Shi et al. [68] presented a new defense strategy termed adversarial supervised contrastive learning, which integrates AT with supervised contrastive learning. This approach aims to improve the resilience of deep neural network (DNN)-based models while maintaining their accuracy on clean data. Furthermore, Shao et al. [69] presented a two-step, effective protection mechanism for textual backdoors known as backdoor defense model based on

detection and reconstruction: (1) recognizing suspicious terms in the sample and (2) recreating the original text by replacing or deleting words. However, given the complexity of cybersecurity, relying solely on these studies is not sufficient. Further research and collaboration are needed to address this pressing issue comprehensively.

In the realms of images and graphs, there has been a focus on the innovation and optimization of defense methods. The primary objective of Li et al. [70] was to create and assess the applicability of SecureNet, a key-based access license approach to shielding the DNN models' IP. SecureNet utilized private keys in model access, employed a key recovery mechanism, and had countermeasures for adversarial and backdoor attacks. Al-Andoli et al. [71] suggested and tested an approach named AEDPL-DL, which stands for the AE detection-based protection layer in DL models to improve the DL models' resilience against adversarial perturbations. The framework also encompasses a protection layer for adversarial examples, as it reduces the reliability, security, and efficiency of DL applications. Unmasking and purification-based methods are fast becoming popular techniques, and since adversarial patches work in different layers of a computer vision system, Yin et al. [72] integrated a modular defense system for easily addressing each layer and aimed to avoid the common problem of image destruction, reduce distribution shifts, and protect these systems against attacks of different kinds, making such systems more secure and trustworthy. Abdel-Basset et al. [73] proposed and assessed the proposed privacy protection-based federated DL (PP-FDL) framework to prevent privacy-related GAN attacks in non-i.i.d. data settings without compromising the classification accuracy in IoT applications such as smart cities. Zhang et al. [74] proposed the realistic-generation and balanced-utility GAN (RBGAN), which is a face deidentification model to address existing issues of privacy preservation and maintain data usefulness through the introduction of disentangled and symmetric-consistency-guided generation parts in the GAN structure. Luo et al. [75] focused on addressing the problem of achieving a natural accuracy–robustness trade-off in federated learning (FL) when dealing with scenarios with skewed label distributions. To accomplish this, generative adversarial networks for federated adversarial training (GANFAT), which demonstrated better outcomes in terms of traffic security against adversarial attacks on the datasets used in the experiments, are suggested.

In addition, to improve the robustness of neural networks against adversarial attacks, attribution guided sharpening (AGS), which uses explainability approaches such as AGS, employs saliency maps derived from a nonrobust model to direct Choi and Hall's sharpening technique, which diminishes noise in input images before classification [76]. Hwang et al. [77] are credited with this approach and introduced AID-Purifier to increase the resilience of adversarially trained networks by refining their inputs. This auxiliary network operates as an extension to an already trained primary classifier and is trained to use binary cross-entropy loss as a discriminator to preserve computational efficiency. The objective of the investigation conducted by Dai et al. [78] was to introduce an effective defense method named deep image prior-driven defense (DIPDefend) in opposition to adversarial examples. By using a DIP generator to match the target/adversarial input, they observed intriguing learning preferences in image reconstruction. Specifically, during the first stage, the main focus was on obtaining robust features that can withstand adversarial perturbations. This was accomplished by incorporating nonrobust features that are susceptible to such perturbations. In addition, they devised an adaptive stopping strategy tailored to diverse images. The goal of the research by Rodríguez-Barroso et al. [79] was to introduce a dynamic federated aggregation operator capable of dynamically excluding adversarial clients. This operator aims to safeguard the global learning model from corruption. Researchers have investigated its usefulness as a protection against adversarial attacks by incorporating a DL classification model into an FL framework. The Fed-EMNIST Digits, Fashion MNIST, and CIFAR-10 picture datasets were used in the evaluation. In addition, the goal of the research outlined by Choi et al. [80] is to present two simple yet effective mitigation techniques (parallelization and brightness modification) to show how to improve the robustness of cutting-edge defense strategies. Moreover, a unique protection strategy that makes use of perceptual hashing is proposed. This technique creates a hash sequence from a query image via the perceptual image hashing strategy known as PIHA. The text presents a plethora of defense methods against adversarial attacks but lacks coherence and conciseness. It suffers from verbosity and repetition, with an overwhelming amount of technical jargon, making it difficult to follow. In addition, critical analysis or comparisons between methods are lacking, leaving the reader unsure of their relative effectiveness or

practicality. A more focused and streamlined presentation would enhance readability and comprehension. To detect adversarial attacks based on queries, the produced hash sequence is then compared with those from earlier questions. Moreover, the goal of the study described by Li et al. [81] was to introduce a new defensive strategy known as robust training (RT). This approach aims to minimize both the robust risk and standard risk simultaneously. Furthermore, the scope of RT was expanded to a semisupervised mode to bolster adversarial robustness. This extension, SRT, proved effective since the robust risk is unrelated to the true label, and the previously p-bounded neighborhood was broadened to encompass various perturbation types. The research outlined by Lu et al. [82] delved into common backdoor attacks within FL, encompassing model replacement, and adaptive backdoor attacks. On the basis of the initiation round, backdoor attacks were divided into convergence-round and early-round attacks. Researchers have proposed two different security strategies: one that uses backdoor neuron activation to address early-round attacks and the other that uses model preaggregation and similarity assessment to identify and remove backdoor models during convergence-round attacks. Li et al. [83] proposed an effective defense strategy by optimizing the kernels of support vector machines (SVMs) with a Gaussian kernel to counter evasion attacks. In addition, Hassanin et al. [84] aimed to develop an attack-agnostic defense method that integrates a defensive feature layer into a well-established DNN architecture. Through this integration, the effects of illegal perturbation samples in the feature space are lessened. Liu and Jin [85] explored the creation of deep neural architectures via a multiobjective evolutionary algorithm that is resistant to five common adversarial attacks. Furthermore, Pestana et al. [86] presented for the first time a class of robust images that are easy to defend against adversarial attacks and that recover better than random images. To improve the adversarial robustness of deep hashing models, the focus has been on investigating semantic-aware adversarial training (SAAT) [87]. Shi et al. [88] proposed an attack-invariant attention feature-based defense model (AIAF-Defense) in an effort to improve the defensive model's capacity for generalization. In the study by Yamany et al. [89], in autonomous vehicle (AV) contexts, a unique optimized quantum-based federated learning (OQFL) framework was created to automatically update hyperparameters in FL via a variety of adversarial techniques. Lee and Han [90] suggested a causal attention graph convolutional network (GCN), and Zha et al. [91] presented a clear and innovative method for performing adversarial steganography while improving the opponents' requirements. In addition, Rodríguez-Barroso et al. [92] discussed robust filtering of one-dimensional outliers (RFOut-1d), a novel federated aggregation operator, as a robust defense against backdoor attacks that poison models. In addition, in the study by Kanwal et al. [93], the aim was to present a feature fusion model that integrates a pretrained network model with manually created features. The objective is to obtain robust and discriminative characteristics. On the other hand, one paper used datasets of text and images; in the investigation by Nair et al. [94], a privacy-preserving framework named Fed_Select was introduced. This framework focuses on ensuring user anonymity in Internet of medical things (IoMT) environments when analyzing large amounts of data under the framework of FL. Fed_Select minimizes potential vulnerabilities during system training by reducing the gradients and participants through alternative minimization. The framework guarantees user anonymity by using hybrid encryption approaches and runs on an edge computing-based architecture. It also has the added benefit of lessening the strain on the central server. These studies present a vast array of defense strategies against adversarial attacks but suffer from several disadvantages. They lack coherence and clarity, overwhelming the reader with technical details. Moreover, these studies may overlook real-world applicability or scalability, hindering their practical utility.

4.2.2 Malware, intrusion, and anomaly detection

This section includes detecting malware, hacks, anomalies, or any threats through adversarial attacks, as this section consists of 7 contributions out of 51.

Numerous studies have been conducted using text datasets. Gungor et al. [95] presented RObust Layered DEFense against adversarial attacks toward IIoT ML-IDMs, which is based on the use of the denoising autoencoder for a better prediction result, as well as protection. In the study by Shaukat et al. [96], ten malware detectors based on neural networks were created. One of these detectors was trained without being exposed to adversarial attacks, but the other nine were trained via a specific adversarial approach. A novel technique is

presented to account for the features of various adversarial attacks and leverage the effectiveness of these detectors against evasion tactics. Using a mix of various adversarial approaches, a neural network is trained in this manner, yielding the best performance out of all 11 detectors. Rathore et al. [97] developed a proactive adversary-aware framework to build Android malware detection models that are more effective when faced with hostile obstacles. In addition, Jia et al. [98] introduced the ERMDS, which uses a wide range of model-agnostic adversarial cases to attempt to provide a more realistic assessment of model performance. These studies have several drawbacks. These methods lack real-world applicability because they focus solely on limited datasets. Furthermore, while some techniques show promise, training methods may not adequately prepare models for diverse adversarial scenarios.

Moreover, Lin et al. [99] answered several important questions: (1) Which adversarial attack function is the most effective at eluding an ML-based network intrusion detection systems (NIDSs)? (2) Which ML algorithm is resilient to hostile attacks? (3) To what extent does the transferability property of an adversarial attack affect the performance of an ML-based NIDS? (4) How can different adversarial assaults be thwarted against an ML-based NIDS? (5) How can the particular ML model that an ML-based NIDS employs be found? However, in the fields of anomaly, intrusion, and malware detection, Xue et al. [100] aimed to provide an early warning system and line of defense against adversarial assaults for electromyography (EMG) signals by recommending a correlation feature on the basis of the Chebyshev distance between neighboring channels. Kopcan et al. [101] proposed systems for identifying anomalies in autonomous transport, encompassing roads, railways, and unmanned aerial vehicles. Two anomaly detection models, namely, an adversarial autoencoder (AAE) and a deep convolutional generative adversarial network (DCGAN), were developed on the basis of the frameworks introduced in Autoencoders (2020) and Deep (2020). The training process utilized image datasets, including the MNIST, Fashion-MNIST, and CIFAR10 datasets. This study has several limitations. They may not fully consider the diverse nature of real-world adversarial attacks or their potential impact on intrusion detection systems. In addition, the proposed solutions might not be sufficiently robust or scalable to handle evolving threats across various domains, potentially limiting their effectiveness in practical applications.

4.2.3 Generative adversarial attack and defense strategy

Within the generative adversarial attack and defense strategy category, 7 of the 51 papers focused on two subcategories: the papers that used text datasets and the papers that used image and graph datasets. These papers explore how to generate an adversarial attack and how to defend against this attack.

In the field of text datasets, two papers were published. Zhan et al. [102] presented MalPatch as a method of adversarial attack against DNN-based malware detection systems. MalPatch produces attack-independent adversarial patches to attack various types of detectors; it injects the patches into malware samples to show the evading results and discusses the countermeasures. In the study by Katebi et al. [103], a DL approach for clustering malware in sequential data was presented, and its vulnerability to adversarial attacks was investigated. The method is applied to Android application data streams via static features from the Drebin, Genome, and Contagio datasets. Postattack, deep clustering algorithms yield an FPR exceeding 60% and an accuracy below 83%. However, implementing the suggested defense method mitigates FPR and enhances accuracy on the basis of the findings. In the study by Zhuo et al. [104], a novel black-box attack approach was introduced, which imposes a stringent constraint on a safety-critical industrial fault classification system. This method limits perturbations to a single variable to create adversarial samples. In addition, variable selection is guided by a Jacobian matrix, which makes hostile samples invisible to the human eye in the dimensionality reduction space. By using AT, the research suggests a matching defense tactic that successfully thwarts one-variable attacks and improves classifier prediction accuracy. These studies present notable shortcomings. While the DL approach shows promise for clustering malware, its susceptibility to adversarial attacks raises concerns about its robustness in real-world scenarios. Despite the proposed defense methods, the postattack performance metrics indicate significant vulnerabilities, highlighting the need for more resilient solutions. Similarly, the introduction of a novel black-box attack underscores potential weaknesses in safety-

critical systems. Although the research suggests effective defense tactics, the study's focus on mitigating one-variable attacks may overlook broader security concerns and fail to address multifaceted adversarial threats.

Conversely, in the domain of image and graph datasets, Husnoo and Anwar [8] sought to develop a one-pixel danger model for an IIoT environment. It introduces an innovative image recovery defense mechanism using the accelerated proximal gradient method to identify and counteract one-pixel attacks. Using the CIFAR10 and MNIST datasets, the experimental results revealed the high efficacy and efficiency of the suggested solution in identifying and thwarting such attacks within advanced neural networks, notably LeNet and ResNet. Zhao et al. [105] created adversarial examples to trick image captioning models to protect private information contained within photos. With five versions, these user-oriented adversarial examples enable users to conceal or change important information in the text output, thus protecting personal information from picture captioning models. Xu et al. [106] created two cutting-edge defensive techniques to address hostile cases in deep diagnostic models: misclassification-aware hostile training (MAAdvT) and multPerturbation adversarial training (MPAdvT). They examined quantitative classification results, intermediate features, feature discriminability, and label correlation for both original and adversarially altered images, delving into the investigation of how adversarial cases affect models. Finally, Soremekun et al. [107] focused on strengthening robust models developed with projected gradient descent (PGD)-based robust optimization by introducing and thwarting backdoor assaults. These studies have several drawbacks. While the proposed defense mechanism shows promise against one-pixel attacks, its applicability to real-world scenarios beyond experimental datasets remains uncertain. In addition, the study may lack comprehensive evaluation across diverse neural network architectures and datasets, potentially limiting its generalizability.

4.2.4 Defense robustness

Within this category, 10 articles of the 51 selected papers were further divided into two distinct subcategories: text datasets and image and graph datasets. In the field of text datasets, Roshan et al. [52] discussed crucial NIDS topics, adversarial attacks, and defense mechanisms for strengthening ML- and DL-based NIDS. Charfeddine et al. [108] offered a detailed understanding of ChatGPT's impact on cybersecurity, privacy, and enterprises concerning various adversarial and protective concepts, including the injection of malicious prompts and NIST security frameworks. First, it suggests secure practices for enterprises; second, it raises some ethical issues; and third, it analyzes potential threats in the near future and the solutions to them. Broadly, Khaleel [109] proposed a method to improve the defense mechanisms of AI learning models used in cybersecurity. On the basis of the Edge-IIoTset dataset for IoT and IIoT applications, techniques such as AT and input preprocessing were incorporated into the study.

In the context of defense robustness, which uses image and graph datasets, the research presented by Meng et al. [110] delves into various classical and cutting-edge adversarial defense strategies within electroencephalogram (EEG)-based brain–computer interfaces (BCIs). In particular, the article evaluates nine defense strategies on two EEG datasets and three convolutional neural networks (CNNs), creating a thorough benchmark to determine each strategy's efficacy in the context of BCIs. To establish a proper framework for systematically solving the facial image data privacy issue, Ul Ghani et al. [111] proposed the use of a privacy-preserving self-attention GAN in conjunction with clustering analysis and a blockchain. The CelebA outperformed the state-of-the-art methods in terms of image realism and privacy preservation for various use cases. Finally, the objectives of the study by Wei et al. [112] were to improve the robustness of DNNs against targeted bit-Flip attacks in security applications. The proposed ALERT defense mechanism incorporates source-target-aware searching and weight random switch strategies while achieving high network accuracy. The primary goal of Shehu et al. [113] was to present LEmo, a cutting-edge technique for classifying emotions that uses facial landmarks and is resistant to hostile attacks and distractions. To compare LEmo with these seven cutting-edge techniques, researchers have compared it with neural networks (ResNet, VGG, and Inception-ResNet), emotion categorization tools (Py-Feat, LightFace, and Adv-Network, DLP-CNN), and anti-attack techniques (Adv-Network, Ad-Network). To assess the robustness of the LEmo approach, three different adversarial attack types and a distractor attack were used. In the study by Nayak et al. [114], to address the problem of giving a

pretrained classifier verifiable robustness assurances within the limitations of a small amount of training data, the authors developed the DE-CROP technique. To train the denoiser, this method consists of two main steps: (1) creating a variety of boundaries and interpolated samples and (2) effectively integrating these created samples with sparse training data. The proposed losses that guarantee feature similarity between the denoised output and clean data at both the instance and distribution levels are used to train the denoiser. Yin et al. [115] aimed to study the practical security risks of the IIoT, investigating transferable adversarial attacks and proposing methods to enhance transferability while providing deployment guidelines against such threats. In addition, in the mentioned study [116], for AT, the authors suggest two approaches called distribution normalization (DN) and margin balance (MB). By standardizing each class's features, the DN guarantees uniform variance in all directions. This normalization aids in removing intraclass directions that are simple to manipulate. On the other hand, MB serves the purpose of equalizing the margins between various classes. By doing so, it becomes more challenging to identify directions associated with smaller margins, thus making it harder to launch attacks on confusing class directions. These studies have certain limitations. While crucial topics in NIDSs, EEG-based BCIs, and defense mechanisms are discussed, an in-depth evaluation or comparison of the proposed strategies may be lacking, potentially hindering their practical applicability. While the LEmo technique shows promise in emotion classification, evaluations against adversarial attacks may lack comprehensive testing across diverse scenarios or datasets and face challenges in ensuring the generalizability of the DE-CROP technique across different datasets or domains because of its reliance on specific training data characteristics. Finally, while the suggested approaches for AT, DN and MB, aim to increase model robustness, their effectiveness in mitigating attacks across diverse datasets or scenarios may require further validation.

4.3 Deep and scientific analysis

This review studies defense strategies and methods for ML adversarial attacks on diverse applications, showing the most efficient and effective choices. The articles of this study have made some relevant suggestions and offered significant ideas, although they may need a few improvements. There is no available systematic comparison or benchmarking of the proposed defense approaches with existing state-of-the-art methods, which makes it difficult to evaluate the relative merits of all approaches comprehensively. Furthermore, different sections of the system have different evaluation metrics, which makes it difficult to directly equate the performance level across different areas. Therefore, this article offers a deeper understanding of the practical applicability and limitations of these methods. Future work may include a broad review of adversarial attack transferability against different defense techniques and the possible trade-offs between model robustness and interpretability.

On the other hand, even though the protection of industrial systems and the level of efficiency of algorithms in the industrial field are important, of the 51 papers that we finalized, we note only four papers [8,94,104,115] that discussed the industrial field. One of the groups was an analysis of how the attack was carried out, the second was how to preserve privacy via FL, and the third was how to preserve privacy via unified learning. Several interesting studies [8] have discussed how to defend against one-variable attacks, and according to these studies, the DNN algorithm was used. However, there are no other studies in this field, and our questions include the following: What of the other attack types? And other types of algorithms in this concern? This area is very expansive, and much work is needed in the future. Moreover, numerous studies have shown that classical FL security is still at risk of privacy attacks caused by data leakage and the opportunity for an adversarial attack during gradient transfer operations. Nevertheless, we have only come across one study concerning this aspect and on FL in the industrial field. Owing to the rise in AI methods, understanding the adversarial attack process is now critically essential. Nevertheless, the 51 papers included in the study did not use or address the explainability of ML algorithms (XAIs) during and after the attack, despite their importance.

Similarly, IoT applications are very critical, and the adversarial attack exposure challenge is very high. However, there are no studies in this respect except for the study that established the connection between

industrial applications and IoT applications [8]. In addition, the implementation of new defense techniques will likely lead to the formation of more computationally complicated and difficult-to-implement components, and thus, they will be disadvantageous with respect to operability and scalability. In addition, the articles could gain quality from a more detailed analysis of the limitations, assumptions, and vulnerabilities of the suggested methods, which would give them a greater picture of the operational potential of the methods in real-life situations. However, the data from these surveys have observable limitations, and as a result, the overall conclusion emerges to improve defense techniques in ML applications.

5 Discussion

This section focuses on three key aspects related to the defense strategies and methods used in adversarial attacks: motivations, challenges, and recommendations.

5.1 Motivations

This section addresses five main topics related to the motivation for defending against adversarial attacks: (1) malware detection and cybersecurity, (2) adversarial defense in specific domains, (3) ML vulnerabilities and AT, (4) privacy and ethical concerns in AI applications, and (5) FL and network intrusion detection. These groups help organize the paragraphs on the basis of common themes and topics related to defenses against adversarial attacks (Figure 12).

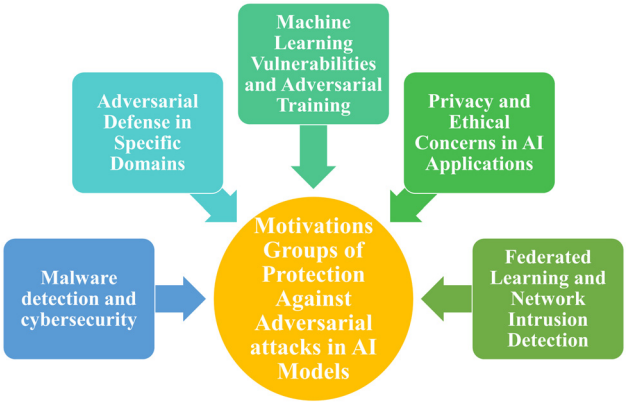

Figure 12: Motivations of protection against adversarial attacks in AI models. Source: Created by the authors.

5.1.1 Malware detection and cybersecurity

In the malware detection and cybersecurity roles, the dynamic evolution of malware and the vulnerability of adversarial attacks against DL-based detectors are discussed, emphasizing the importance of addressing these vulnerabilities. Numerous investigations have been carried out on this topic. Given the increasing concerns surrounding cybersecurity, the continuous evolution of malware remains a significant challenge. DL-based malware detectors are considered promising solutions, but their susceptibility to adversarial attacks underscores the importance of thorough validation. It is imperative to ensure the resilience of these detectors against such attacks [96]. In addition, AID-Purifier, introduced by Hwang et al. [77], is motivated by its role as a lightweight auxiliary network designed to purify adversarial examples. Notably, it introduces a unique

purification approach using a straightforward discriminator, which distinguishes itself from previous purifiers by allowing purified images to exist in out-of-distribution regions. Antimalware learning (AML) is a new area of study that is crucial for understanding and defending against adversarial attacks to protect computer networks from various cybersecurity risks [52]. In the contemporary landscape, billions of users rely on Android smartphones, rendering them appealing targets for malware designers seeking lucrative opportunities [97].

5.1.2 Adversarial defense in specific domains

With respect to adversarial defenses in specific domains, the focus is on tailored defense strategies for specific areas, such as image recognition and industrial fault classification systems, which recognize the need for domain-specific protective measures. According to Meng et al. [110], their work represents a pioneering effort in the field of adversarial defense for EEG-based BCIs, which holds significant importance for the practical implementation of BCIs. Furthermore, by comparing nine contemporary and traditional defense strategies across three CNN models and two EEG datasets with different assault scenarios, they created a standard for adversarial protection in BCIs. Similarly, attribution guided sharpening (AGS) creatively uses attribution values produced by a nonrobust classifier to direct the adversarial noise denoising process, as proposed by Perez Tobia et al. [76]. This approach serves as a novel defense strategy for nonrobust models without the need for additional training. Research on emotion categorization has gained significance because of the proliferation of intelligent systems, including human–robot interactions. DL models, as highlighted in the study by Shehu et al. [113], have excelled in various classification tasks. However, their vulnerability to diverse attacks stems from their homogeneous representation of knowledge. Inspired by the ability of deep image priors to capture extensive image statistics from a single image, a robust defense method against adversarial examples is needed [78]. Furthermore, Zhuo et al. [104] were motivated by the absence, to date, of proposed and analyzed adversarial attack and defense methods specifically tailored for industrial fault classification systems. The rationale behind [112] was to meet the demand for large-scale usage of DNNs in security-sensitive domains that include self-driving cars and health care.

5.1.3 ML vulnerabilities and AT

The following sections discuss ML models' susceptibility to adversarial attacks and how AT can be used to make ML models more resilient in the context of ML vulnerabilities and AT. The increasing dependence on ML within code-driven systems for smart devices and applications has become a crucial aspect of human reliance. While ML has been the subject of much research to address real-world problems such as image categorization and medical diagnosis, a significant gap exists in the understanding of adversarial attacks on safety-critical networked systems. This is critical because attackers can exploit adversarial samples to circumvent pretrained systems, resulting in heightened false positives in black-box attacks and finding patterns in data that diverge from a model of typical behavior, which is known as anomaly detection. Robust anomaly detection systems are essential across various domains [101]. The purpose of the study by Khaleel [109] was to contribute to filling this notable blind spot with respect to susceptibility to adversarial attacks in ML for cybersecurity use cases in the literature.

With the growing autonomy of vehicles on roads, railways, and unmanned aerial vehicles, there is an increasing need to address anomalous situations not covered by trained models, thereby increasing safety risks. Researchers have proposed AT, which involves adding adversarial samples to the training data, to improve the adversarial robustness of neural networks. This strategy, however, might cause overfitting to certain adversarial approaches, which would lower standard accuracy on clean images. In addition, there has been a notable surge in interest, both among academics and industry professionals, in defending DL systems against one-pixel attacks and guaranteeing their robustness and endurance [8]. AT has emerged as a potent defense method for minimizing adversarial risk, and accurate predictions for both benign examples and their

perturbed counterparts within the ball have been proposed. The motivation behind the study by Li et al. [81] aimed to specifically and cooperatively increase adversarial robustness and accuracy.

As the complexity and abundance of malware threats continue to rise, traditional signature-based detection methods are experiencing diminished effectiveness [98]. Despite the commendable strides made in industrial applications through rapid ML development, these achievements coexist with notable security vulnerabilities [83]. Adversarial attacks, particularly data poisoning-enabled perturbation attacks in which false data are injected into models, profoundly affect the process of learning and degrade the accuracy and convergence rates without benefiting deeper networks [84]. While progress has been made in developing robust architectures, a common weakness persists in existing AT approaches, which focus on a singular type of attack during evaluation [85]. Even while DL offers state-of-the-art image detection solutions, vulnerabilities persist even against minor perturbations [86]. Deep hashing models are vulnerable to security threats when hostile cases are recognized, highlighting their susceptibility [87]. Current research highlights how ML algorithms are susceptible to transfer-based attacks in real-world black-box situations [115]. It is important to examine the resilience strategy that uses PGD as a universal first-order adversary, especially in light of how it behaves when facing attacks that are fundamentally different from backdoors [107]. Adversarial attacks, characterized by iterative sample movement, emphasize the importance of traversing decision boundaries for classification loss ascent [116]. The demonstrated vulnerabilities in GCNs further underscore the need for enhanced defense mechanisms [90]. In the pursuit of adversarial optimization, insufficient attention has been given to exploring the collaboration between cover enhancement and distortion adjustment, leading to potential local mode collapse [91]. Recognized as an effective strategy, AT involves considering benign and adversarial examples together in the training stage to bolster the robustness of DNNs [68]. The intricate interplay between adversarial attacks and model misbehavior through backdoors becomes apparent in the context of infected models performing well on benign testing samples [69]. In the domain of person reidentification (RE-ID) adversarial attacks, introducing noise or foggy material in images disrupts the model's ability to recognize the similarity between gallery and probe images, resulting in drastic changes in recognition outcomes [93].

5.1.4 Privacy and ethical concerns in AI applications

The section highlights concerns about privacy breaches and ethical considerations in AI, particularly in image labeling, as well as potential misuse of AI in advertising, underscoring the ethical implications of AI applications. With respect to shifting focus, two studies explored privacy and ethical concerns in AI applications. The significance of the study by Katebi et al. [103] becomes evident in light of the escalating data streams within Android systems, underscoring the necessity for a fundamental analysis. The temporal nature of streaming data, which are available for a specific duration, poses a challenge because of the evolving data models. The proposed clustering methods should effectively handle this temporal characteristic.

Furthermore, adversaries creating malware strive to emulate benign samples, creating a challenge for precise clustering. To address this issue, static analysis of Android malware involves evaluating distinct features between benign and malicious samples. AI systems, which are utilized for the automatic labeling of images on social networks, hold potential for positive applications, such as providing text descriptions for visually impaired individuals. However, concerns arise with the use of cross-modal techniques, raising issues about potential unethical purposes, such as analyzing personal information for advertising. This emphasizes the crucial need to prevent privacy breaches [105]. Owing to the large-scale adoption of DNNs, the threat of privacy attacks against DNN models is always increasing, and hence, intellectual property (IP) protection is needed for such models [70]. The basis for the study by Zhang et al. [74] was the increase in the use of computer vision and surveillance technologies, which threatened privacy on the basis of facial identity information. This spurred the creation of the RBGAN, with the intent of providing optimal privacy while being able to maintain high data utility through the generation of realistic images that preserve attributes.

5.1.5 FL and network intrusion detection

FL and network intrusion detection are examined in a set of papers addressing challenges and vulnerabilities in FL, which encompasses challenges associated with hyperparameter optimization and the vulnerability of NIDSs to adversarial attacks, underscoring the necessity for improved security measures. In this domain, a multitude of articles predominantly rely on techniques such as differential privacy, secure multiparty computation, and homomorphic encryption to safeguard the security of models and data. Nevertheless, these approaches frequently demonstrate inefficiency, particularly in sensitive domains such as healthcare or finance [94]. This inefficiency gives rise to concerns about the suitability of the current literature for application in IoMT systems. Concerns over user privacy are further heightened by FL servers' sincere yet inquisitive attitudes. Establishing hyperparameter settings is crucial for FL performance efficiency, and the automated refinement of these parameters has the potential to contribute to the creation of reliable FL models [89]. As a distributed ML paradigm, FL faces susceptibility to various adversarial attacks because of its distributed nature and the limited access of the central server to data [92]. While NIDSs have embraced ML for detecting a broad spectrum of attack variants, the inherent vulnerability of ML to adversarial attacks poses a risk, potentially compromising ML accuracy in the process [99].

5.2 Challenges

In every research work, there are some challenges and obstacles facing researchers. This section discusses these difficulties in four groups (Figure 13).

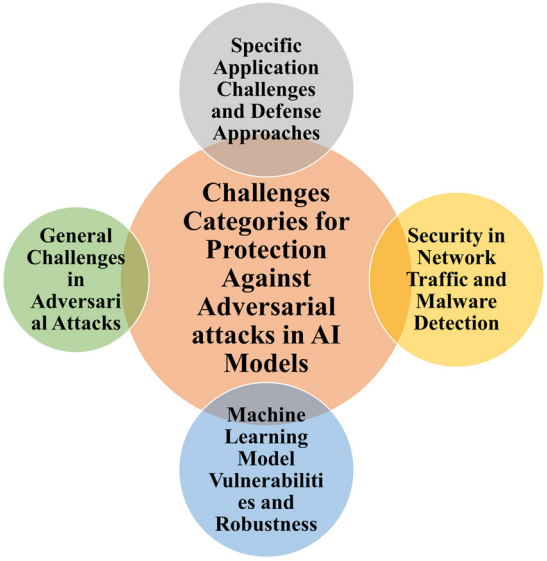

Figure 13: Challenge groups of protection against adversarial attacks in AI models. Source: Created by the authors.

5.2.1 General challenges in adversarial attacks

This section examines the general challenges of adversarial attacks on ML models, including their susceptibility to static malware, time complexity, and need for substantial processing power. A significant concern in the analysis discussed in the study by Katebi et al. [103] was the considerable time complexity, especially when handling high-dimensional data, which results in longer processing times. In addition, similar to previous DL

techniques, this approach required a significant amount of computing power, which was not readily accessible on Android operating system host devices at that time. FL, operating across independent devices with heterogeneous and unbalanced data distributions, faces heightened vulnerability to adversarial attacks, particularly backdoor attacks [82]. In addition, FL, an adversarial attack known as Byzantine poisoning, can be used against a decentralized training strategy that is carried out locally on devices without direct access to training data. It is difficult to handle the defense against these threats in an efficient manner given the inadequacy of existing defenses, as highlighted by Rodríguez-Barroso et al. [79]. Furthermore, a noteworthy difficulty lies in the efficient extraction of dependable semantic representations for deep hashing, impeding progress in adversarial learning and hindering its improvement [87]. The reason for prioritizing the technique of creating highlighted adversarial examples in a black-box scenario is grounded in its noteworthy success rate and practical feasibility, as highlighted by Choi et al. [80]. Even with its broad application, conventional FL is still vulnerable to adversarial attacks during gradient transfer procedures and data leakage [94]. In contemporary society, individuals are heavily reliant on social networks such as Facebook and WeChat, which have profoundly transformed our lifestyles. When users engage in activities such as giving thumbs-up, sharing, and commenting, their personal information becomes susceptible to threats [105].

5.2.2 Specific application challenges and defense approaches

This group focuses on specific application challenges, such as issues in fault classification accuracy. Numerous investigations have been carried out on specific application challenges and defense approaches. The primary defense against malware attacks is provided by antimalware/antimalware software products. However, the current literature indicates that prevailing malware detection mechanisms, such as signature and heuristic methods, struggle to address contemporary malware challenges [97]. In addition, a dataset with a variety of traits not observed during LB-MDS training is the primary obstacle to evaluating the robustness of LB-MDS [98]. They reported that as the value of the kernel parameter increases, hostile samples become increasingly undetectable in their thorough testing on three datasets [83]. Furthermore, prevailing defense methods offer resilience against specific attacks, but developing a robust defense strategy against unknown adversarial examples remains a formidable task [88]. While it does not directly control the feature space during training, AT is acknowledged as the most efficient method for improving adversarial robustness [116]. Adversarial steganography has shown state-of-the-art results in the sequential min–max steganographic game. Its goals are to fool steganalysis algorithms and improve embedding security. However, the existing approaches suffer from high computing costs and limited convergence [91]. Addressing defense against model-poisoning backdoor attacks, a significant challenge in FL, was addressed in a previous study [92].

Moreover, it was noted in another study that, compared with the IMDB dataset, the average sentence length of the SST2 dataset is significantly shorter, making the addition of a trigger to the SST-2 dataset more prone to causing significant disruptions and compromising sentence naturalness [69]. Recently, subtle changes in clean images, known as adversarial examples, have demonstrated the high susceptibility of DNNs. Several strong defensive approaches, such as ComDefend, address this problem by focusing on correcting adversarial samples with well-trained models that are derived from large training datasets that contain pairs of adversarial and clean images. These methods, however, ignore the wealth of internal priors included in the input images themselves and instead mostly rely on external priors acquired from massive external training datasets. This constraint hinders defensive model generalization against adversary cases with skewed image statistics relative to the external training dataset [78]. The challenge that forms the theme of the study by Yin et al. [72] was to develop defense mechanisms against adversarial patches, which deceive DL models by applying stickers or particular patterns on objects. This vulnerability was highly dangerous and affected security-critical domains such as security vision systems and self-driving, which we are familiar with. Finally, the difficulty in reidentification (RE-ID) stems from the uncertain bounding boxes of individuals; significant variations in brightness, position, background clutter, and obstruction; and the ambiguous presence of graphics [93].

5.2.3 Security in network traffic and malware detection

This topic addresses security in network traffic and malware detection, highlighting challenges in intrusion detection. Given the difficulty of training generative models and their subpar performance in scenarios with limited data, it becomes crucial to adopt an approach, as suggested by Nayak et al. [114], that generates additional data to mitigate overfitting. The study by Yamany et al. [89] highlights the potential of FL for data privacy but underscores its vulnerability to adversarial attacks, with a specific emphasis on data poisoning attacks that entail injecting malicious vectors during the training phase. In addition, the effective establishment of a resilient FL model against adversarial attacks relies significantly on the appropriate tuning of hyperparameters. In the study by Soremekun et al. [107], the main task was to make it easier for software to automatically identify insecure (backdoor-infected) ML components. The weakness of ML-based intrusion detection systems in the IIoT environment necessitates robust defensive strategies [95]. The matter discussed in the study by Charfeddine et al. [108] was related to the steps transferring in the ethical and logistical aspects of implementing ChatGPT in cyber protection. Considering the possibilities of the proposed solution for improving security activities alongside threatful applications, attention to the ethical use of applications, as well as the problem of effective defense within the NIST framework, was discussed. The problem stated by Ul Ghani et al. [111] consisted of meeting strict privacy constraints when working with facial image data due to the increasing use of facial recognition technology.

5.2.4 ML model vulnerabilities and robustness

This group explores ML model vulnerabilities and robustness, covering issues such as data-driven algorithm vulnerabilities, adversarial security concerns, challenges in adversarial steganography, and problems in NLP tasks. In the study by Hassanin et al. [84], incorporating adversarial examples during classifier training presented two key challenges. First, the features extracted from the input instances, which are employed in the classification step, must be robustly guarded against various adversarial attacks. This guarantees that features from both benign and adversarial instances belonging to the same class are closely aligned in the feature space. Second, achieving a clear separation of classification boundaries between different classes becomes imperative. According to Liu and Jin [85], by predicting the specific types of attacks that a ML model might encounter is practically impossible. In addition, Yin et al. [115] identified security vulnerabilities in existing ML models, where the inclusion of carefully crafted perturbations to input samples could lead to erroneous decisions.

Moreover, attacks causing graph distortion introduce bias that misguides model predictions; addressing this bias is crucial for robust GCNs [90]. Furthermore, Lin et al. [99] outlined three objectives: (1) to assess several protection strategies against adversarial attack functions to ascertain which one is the most successful for NIDSs. (2) Considering the scarcity of publicly available datasets with adversarial samples, we create an adversarial dataset for NIDSs. (3) To forecast an ML-based NIDS model on the basis of the outcomes of an adversary assault, the utility and challenges of identifying the underlying ML method are highlighted. This challenge stems from the diverse ML algorithms used in creating ML-based NIDSs. The use of DL has yielded excellent results in different areas. However, DL models face major difficulties in identification and protection against adversarial samples (AEs) [71].

As indicated in the study by Meng et al. [110], the defense tactics covered are not exclusive of one another; a defense strategy may combine several ideas. Nevertheless, many defense strategies are vulnerable to fresh attacks and can only repel certain kinds of attacks. Furthermore, since black-box assaults involve less knowledge of the target model than white-box attacks do, protecting against them is typically easier. Despite significant strides in DL over the past decade, susceptibility to adversarial attacks has persisted. These attacks can cause neural networks to anticipate things incorrectly because they resemble clean data. Furthermore, DL models frequently serve as “black boxes,” lacking explanations for their outputs, as noted in the study by Perez Tobia et al. [76]. Even though one-pixel attacks are unnoticeable to the human eye, they can have a substantial influence on DNN accuracy. Such attacks can have severe consequences in important fields such as AVs and

healthcare [8]. DNN models that incorporate many data patterns have significantly increased the accuracy of fault classification. Nevertheless, these models, which are based on data, are vulnerable to adversarial attacks, and minute variations in the samples might result in imprecise fault predictions. Finally, Shi et al. [68] highlighted that DNNs might suffer severe performance degradation from attacks utilizing adversarial samples.

5.3 Recommendations

This section discusses future directions and recommendations for researchers in the field of defense against adversarial attacks (Figure 14).

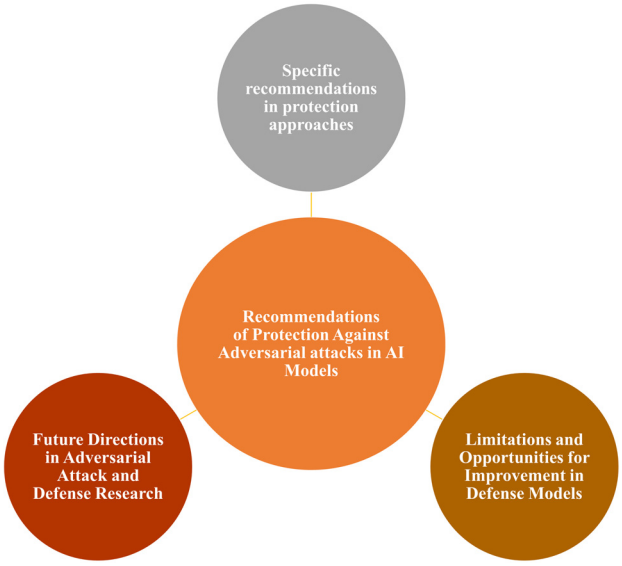

Figure 14: Recommendation categories of protection against adversarial attacks in AI models. Source: Created by the authors.

5.3.1 Future directions in adversarial attack and defense research

This section discusses potential future research directions for adversarial attack and defense. Several studies have focused on future directions in adversarial attack and defense research. However, Shaukat et al. [96] covered various adversarial attacks, exploring the robustness of evasion techniques beyond the specified ten remains an open research direction. In addition, a future focus should include assessing false-positive and false-negative rates by poisoning benign samples. Future advancements in the study by Katebi et al. [103] are anticipated to improve generative adversarial network (GAN)-based attack and defense techniques . It is expected that using GAN methods as a data-generating methodology will produce remarkable results, making them useful weapons for both attack and defense. Several suggestions for additional studies are listed in the study by Meng et al. [110]. These include the following: (1) Optimizing resilient training to increase model accuracy on adversarial instances while maintaining maximum accuracy on regular cases. (2) Improving reliable extrapolation from unknown hostile data is a crucial aspect of RT. Given the variability in EEGs, robust generalization is especially vital for BCIs. (3) Developing input transformations specific to EEGs to concurrently enhance both accuracy and robustness. (4) Investigating defensive strategies for traditional ML models in the context of BCIs.

In addition, the scientists in the study by Lu et al. [82] investigated common backdoor attacks in FL, such as model replacement and adaptive backdoor attacks. They recommended delving deeper into the Byzantine

backdoor attack in FL, which involves crafting multiple malicious models simultaneously to elude defense mechanisms. The proposed solution and algorithm operate independently, necessitating image recovery before DNN classification. The author proposes a potential avenue for future research: developing a DNN layer that integrates this algorithm as a preprocessing step before classification. The incorporation of this layer into other cutting-edge DNN architectures has the potential to bolster robustness and resilience. In the study by Rathore et al. [97], examining collusion-based adversarial attacks against malware detection methods was the main goal. Studying the adversarial robustness of different features, such as hybrid malware detection models and clustering, is another aspect of exploration. To preserve an adversarial advantage in malware detection, the proposal also entails exploring a game-theoretic method. Future directions for Zhan et al. [102] include increasing the applicability of MalPatch in that it should be able to inject patches throughout malware while maintaining its malicious functionality to improve its evasion capability. In addition, improvements in black-box attacks through the use of DRL and GANs, among other methods, are important. Furthermore, GANFAT is proposed for other non-IID environment settings, such as feature skew, quality skew, and quantity skew. Continuous work will be performed to gradually improve defense approaches that decrease the applicability of new unknown attacks. Future studies will also analyze more stable aggregation techniques that work to maintain and allow for fair decision making when assessing different models [75].

Moreover, Yamany et al. [89] focused on a substantial chance to work with large-scale heterogeneous data types and control further FL hyperparameters, such as batch sizes and the number of participating clients. Investigating the proposed framework's resistance to various adversarial attack forms, such as inversion attacks, is another possibility; Zha et al. [91] suggested several directions for additional study in this field. First, as long as the distortion-minimizing embedding framework is relevant, it is still feasible to extend the suggested concepts to other data domains, including the JPEG image domain, the unnatural image domain, or even the audio domain. Second, adding the paired loss to GAN-based steganography techniques could enhance their overall robustness and performance. Third, the suggested adversarial steganography approach may determine the payload capacity for each image by eliminating the assumed payload rate constraint for each image. This approach can help steganographers address the payload allocation issue in batch steganography scenarios. Furthermore, Lin et al. [99] noticed many studies and proposals on different learning models, attack strategies, and defense tactics that are available in the literature. However, addressing every possible combination of these components in a single study is difficult. Future research should investigate broadening the scope of diversity by incorporating more schemes. Finally, the future work outlined in the study by Xue et al. [100] seeks to provide more potent defensive and detection strategies against hostile assaults. For future work, the study by Abdel-Bassetet al. [73] aims to explore blockchain technology for constructing safe multiparty IoT applications in smart cities that are shielded from privacy invasion attacks. In addition, it aims to examine the feasibility of DL in offloading IoT services in extensive networks of the IoT, with cooperating fog nodes and cloud servers.

5.3.2 Specific recommendations in protection approaches

This topic focuses on challenges and recommendations in current defense approaches, including suggestions for evaluating robustness, exploring defensive strategies for traditional ML models, and addressing uncertainties in the performance of proposed methods in different contexts. Perez Tobia et al. [76] advocated for the exploration of new attacks to challenge their defense, emphasizing that this process is crucial for identifying weaknesses and advancing more reliable ML. They demand more work at the nexus of robustness and explainability. This article also recommends extending the examination of attribution-guided denoising (AGD) performance to new domains by examining alternative AGD defenses with different attribution methodologies or denoising methods. The utilization of sophisticated training algorithms, such as the Levenberg–Marquardt algorithm and particle swarm optimization, has the potential to improve classifiers, thereby minimizing misclassifications in the presence of adversarial challenges. Xu et al. [106] recommended examining how well the certified/provable robustness approach would perform to separate training samples that were correctly classified from those that were not. In addition, they recommended investigating the potential improvements associated with this differentiation of training examples.

The AIAF-defense method proposed by Shi et al. [88] possesses certain limitations. For example, its defense efficacy diminishes under high-intensity attacks or intricate target models. In addition, for adversarial detection, the characteristics extracted via specific perturbation techniques might be employed, a fact that warrants further exploration in future studies. Furthermore, Lee and Han [90] distinguished itself as the first study to employ a causal mechanism in proposing a robust model against attacks. While a simple intervention is employed, future research could explore a broader array of methodologies applicable to GCNs. In addition, Shao et al. [69] indicated that they plan to expand the scope of their defensive model to include other NLP functions, such as machine translation, semantic matching, and reading comprehension. The goal of the study outlined by Roshan et al. [52] was to examine how the proposed method impacts various ML and DL architectures. In addition, researchers have suggested extending this approach to explore the transferability concept within adversarial ML. They also recommend applying the same method to investigate concept drift in network streaming data-based NIDSs. Finally, Kanwal et al. [93] suggested that solutions for various adversarial attacks impacting person reidentification (RE-ID) performance could be devised. By employing data augmentation techniques to increase the quantity of input photos in datasets, the efficiency of the suggested model may be further improved.

5.3.3 Limitations and opportunities for improvement in defense models

This section highlights limitations and opportunities for improvement in existing defense models, such as challenges in evaluating the robustness of specific methods, time-consuming aspects of certain approaches, and the potential for enhancing defense models through further investigation and adaptation to different scenarios. Conducting a comprehensive evaluation of the robustness of MDS poses a considerable challenge. The findings of Hwang et al. [77] improved the performance of adversarially trained networks over time and functioned well with other adversarial purifiers, such as NRP and PixelDefend. It is still debated whether adversarially trained networks should always have one or more adversarial purifiers present. In the study by Jia et al. [98], the main approach used three different kinds of obfuscation spaces to measure how well the current MDS performs against adversarial attacks. This research confirms that obfuscation methods are useful for evaluating MDS resilience. Nonetheless, hostile sample production goes beyond obfuscation strategies. While the LEmo method proposed by Shehu et al. [113] demonstrated robust generalization across posed datasets such as CK + and KDEF, which is commendable, uncertainties persist regarding its performance on nonposed datasets. Future research should therefore test this technique with nonposed databases to ascertain its applicability in broader contexts.

Moreover, during the evolutionary search process, a significant amount of effort must be invested in training each child's architecture to determine its fitness values. In addition, the same methods used in DARTS are adopted in this work, which may limit the overall resilience of the identified structures. Cheng et al. [116] suggested alternative forms of loss functions to achieve the goal of DN and MB, potentially enhancing adversarial robustness. In addition, the assumption that the distribution of each class adheres to a Gaussian distribution for analysis convenience could be replaced with a more intricate distribution for a deeper understanding of adversarial robustness. The proposed enhanced FL methodology in the study by Nair et al. [94] leverages an edge computing-based architecture to enhance gradient privacy and preserve user identity. Rigorous experiments on standard datasets validate the scheme's superior performance in terms of privacy enhancement without compromising other system metrics. Future studies could explore further decentralization of the system architecture and introduce multilevel optimization techniques for enhanced performance. Finally, another avenue for research pertains to corruption robustness, a factor that is likely prevalent in practical scenarios.

6 Gaps, open issues, and innovative key solutions

This section aims to identify gaps in the field for future studies, with the potential to benefit researchers. Each subsection focuses on a specific gap and highlights areas lacking in defense in the context of the adversarial

attack. The following subsections present noteworthy tables, and analyses that provide an overview of the latest advancements in protection against adversarial attacks in AI models found in the literature.

The exploration of protection strategies and methods within the realm of AI models has garnered considerable attention in the realm of scientific inquiry. This subject remains contemporary, with a burgeoning body of research indicating the ongoing expansion of knowledge in this area. Figure 15 illustrates the upward trajectory of research contributions, highlighting the evolution of interest from a mere 3 in 2020 to a noteworthy 17 by the year 2023. This escalating trend underscores the novelty and significance of the topic. Remarkably, the absence of research contributions in 2019 within the specific scope of our investigation accentuated the recent surge in scholarly activity dedicated to understanding and advancing protection strategies and methodologies for AI models in adversarial attacks.

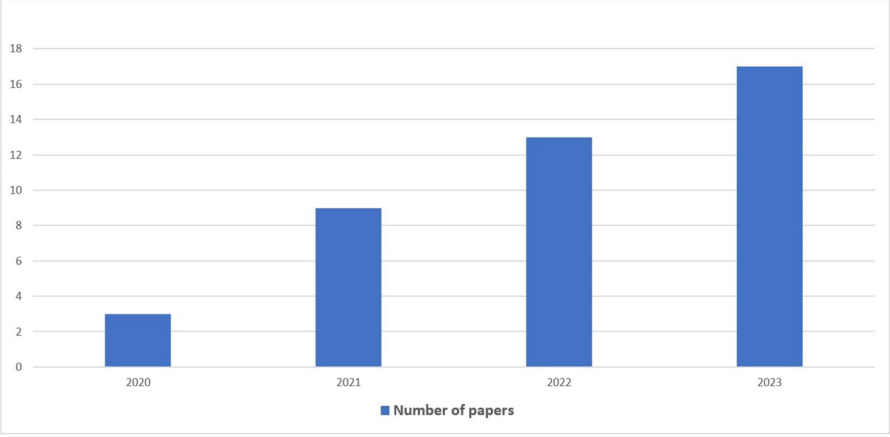

Figure 15: Number of papers on protection against adversarial attacks in AI models. Source: Created by the authors.

As this topic occupies a special place in the research discussion, we have outlined some directions for further research and revealed some innovative solutions.

Gaps and open issues:

- Insufficient collaboration and comprehensive research: With respect to the corresponding current approaches to disruption for adversarial defense methods, what are identified here are insufficiently systemic approaches, and they do not unify meaningfully enough to encompass all forms of cybersecurity. Research cooperation and collaboration become very important because congruent and coordinated research activities are mandatory for creating holistic and consistent defense policies.
- Lack of real-world applicability and scalability: What appears quite frequently is that the potential defenses that are sought can be tested in limited and often quite complicated ways to determine their effectiveness and feasibility. However, there is a need to perform evidential implementation of such solutions in various installations because of the high risk attached to innovative solutions.
- Comparative analysis and benchmarking deficiencies: Defense scenarios are somehow deficient in elaborate critical discussions and benchmarking with other similar countermeasures employed in other counter ending projects. If benchmark measurements had been defined and research and studies for the purpose of comparison had been carried out, then there are chances that there would have been more clarity regarding the efficiency and utility of various approaches.
- Malware, intrusion, and anomaly detection systems: The current research is still somewhat limited in terms of its real-world relevance because datasets employed in experiments tend to be much smaller. Moreover, adversarial examples may also affect IDSs, or the current approaches may not address the various types of adversarial examples for models. Nevertheless, it is critical to recognize that some of the devised solutions might have limited capacity to be effective as the threats progress.

- Defense robustness: Some works may fail to conduct or neglect a detailed examination or include strategies that are recommended, which may ultimately diminish their active application in reality-based environments. This means that techniques may not readily transfer across distributions or across two or more datasets.

Some innovative key solutions:

- Integrated AT techniques: In addition to AT, other related methods, such as supervised contrastive learning, provide a better balance between robustness and accuracy since none of the standard approaches provide a reliable method of achieving both of them simultaneously.
- Two-step textual defense mechanisms: Applying two-step approaches, both for threat recognition and for threat prevention, such as the identification of problematic terms and the reconstruction of text, would also be highly helpful for guaranteeing the security of textual data models.
- Explainability-driven image defense methods: Techniques that attempt to improve the quality of images before classification can serve an opponent in how to manipulate an input to obtain the event in an NN that is of interest, thus making the NN more proficient in adversarial attack defenses.
- Dynamic FL protections: Static possession, logical possession, and general implementation of dynamic aggregation operators in FL may also repair corrupt client influences and exclude malicious clients and thus, overall, correct learning models that are immune to corruption in distributed training processes, such as security performance.
- Semantic-aware adversarial training: Infusing the notions of semantic awareness into AT decreases the vulnerability and enhances the firmness of the models to the overall adversarial noise, and defended models will be able to maintain high levels of accuracy and discriminant power.
- Privacy-preserving FL frameworks: Further work on establishing trustworthy FL as a result of creating the necessary frameworks that will confirm the identity of the user and the possible threats in the training phase of this system, preventing the overload of server systems during the work.
- The focus should be on the development of stable defense measures that are likely to hold sufficient capacity for countering various types of attacks. Training and testing with greater and, consequently, more diverse data can enrich the reality of these models. Furthermore, the use of adaptive learning techniques to fit the new threat adds more weight and overtones to these solutions.
- It is necessary to focus on determining the strategies that can withstand various types of attacks. This includes enhancing the defense strategies applied to DL and contributing to the extent of the attack types. Furthermore, the creation of suppositional settings that reflect real-life adversarial situations may assist in building a model for actual environmental practices.
- More extensive research must be conducted to investigate the overall efficiency of defense mechanisms for adversarial examples for different datasets and the utilization of various neural networks. Hence, it will be possible to enhance the overall robustness by implementing proactive defenses that adapt to the type of threat.

Practical guidance for industry practitioners and policymakers:

- Adopt standardized evaluation frameworks:

  It is difficult to follow a proper comparison pattern or, in fact, have a criterion to determine the effectiveness of various strategies in the academic field of defense.

  Recommendation: Establish clear principles and strategies for safeguarding AI systems and goods to facilitate the possibility of comparing the efficiency of the protected measures across the many applications of AI. This involves the development of bigger datasets that have multiple types of adversarial examples as well as the same for accuracy, robustness, precision, and recall.
- Stressing on the explainability and the transparency:

  In light of this, it is of paramount importance to get a glimpse into the said process with specific focus given to the fact that attacks majorly tend to focus on the models' decision-making abilities.

  Recommendation: Integrating the new generated called explainable AI (XAI) methodology into the defensive approaches. This involves developing methods for the representation of the impacts that

rival attacks inflict on the conclusions of the model that will instill confidence and assist in identifying the flaws.

- Further research on AT strategies should focus on the enhancement of the mentioned AT approaches:

  This article has identified that AT is among the most effective methods that can be used to improve the model's resistance to adversarial attacks.

  Recommendation: Feed adversarial samples into the training process to increase the robustness of the up-model programs. If analysts and other practitioners of the particular field of application of the classifier increase its resistance to adversarial examples but do not consider overemphasis on the enhancement of precision, they will be more interested in the generalization ability of the model.
- Increase the criterion of applicability and workability of the solutions:

  All of these approaches have to be realistic and have a dose that actually can be practically put into practice in match situations.

  Recommendation: Introduce and test out novel distinctive approaches to defense in such sensitive spheres as health care, banking, and auto pilot systems. Check the applicability of these solutions in real-world conditions for practicality while also studying the scalability of the found solutions.
- Encourage interdisciplinary collaboration:

  Adversarial threats are complex and thus to manage them, professionals from diverse fields are required to be sourced.

  Recommendation: Strengthen collaboration between academic institutions and the industry and government for constructing sound defense strategies. This comprises incorporating information acquired on cyber security, AI, data science, and issues of moral significance.
- Meet ethical and privacy issues:

  AI defenses raise very practical ethical questions and the top of them is the question of privacy.

  Recommendation: The development of codes of conduct and regulation independent of the size of the implications that AI security measures' consequences have on users' privacy and data. Thus, it is ensured that policymakers have to ensure that the specific AI technologies are being used appropriately and ethically.

6.1 Dataset availability for defense in adversarial attacks

The significance of the dataset in the context of adversarial attacks lies in its foundational role in training AI models. Specifically, within adversarial attack scenarios, the dataset becomes essential for comprehending vulnerabilities and potential threats. The absence of detailed information about datasets, as observed in the literature analysis, impedes the evaluation of model suitability and generalizability. It is crucial to have precise knowledge about the data type, size, composition, and specifics related to training and testing datasets to assess the resilience of AI models against adversarial attacks (Table 1). Furthermore, the importance of the training data's validity is underscored, where validation ensures an accurate representation of real-world scenarios, and homogeneity ensures consistency for successful generalization across various situations. Transparent disclosure of the dataset's source, whether publicly available or privately collected, is pivotal to ensure the reproducibility and credibility of research findings in the domain of adversarial attack.

6.2 Defense in adversarial attack-based ML/DL techniques

ML techniques not only offer considerable advantages but also present challenges. ML and DL can effectively handle tasks such as image classification, object recognition, and natural language data processing, resulting in advancements in data representation [4]. Transferring data from models trained with one dataset to another is a fascinating topic associated with ML [117]. Nonetheless, a primary challenge lies in acquiring substantial

Table 1: Dataset of defense strategies and methods in adversarial attack

Ref.	Dataset name	Description	Size	Link of the dataset (if it is available)	Is the legally collected dataset?	Public/ Private
[96]	– VirusShare – VXHeaven	The datasets contain Portable Executable (PE) format using malicious files from each malware repository and benign PE files	– 19,000 – 19,000	– VirusShare.com – vx.netlux.org	√	Public
[103]	– Drebin – Contagio – Genome	– The Drebin dataset comprises diverse Android malware families, with samples obtained through the Mobile Sandbox, encompassing applications from 179 malicious families – The Contagio dataset includes both benign and malicious samples collected from mobile devices – The Genome dataset, curated by researchers at the National Science Foundation in the United States, consists of malicious applications	– 5,721 – 28,760 – 1,200	– https://www.sec.cs.tu-bs.de/danarp/drebin/ – N/A – N/A	√	Private
[110]	– Feedback error-related negativity (ERN) – Motor imagery (MI)	– The data, which came from 26 participants, was split into two groups: a test set (10 subjects) and a	– 8,840 – 1,296	– https://www.kaggle.com/competitions/inria-bci-challenge – https://www.bbci.de/competition/iv/	√	Public

(Continued)

Table 1: *Continued*

Ref.	Dataset name	Description	Size	Link of the dataset (if it is available)	Is the legally collected dataset?	Public/ Private
		training set (16 subjects) – Collected from nine individuals engaged in four motor imagery tasks (left hand, right hand, feet, and tongue), the dataset comprised 144 trials per task from two separate sessions				
[76]	– MNIST	– The dataset comprises	– 60,000	– N/A	√	Public
[116]	– CIFAR-10	grayscale images of	– 60,000	– https://www.cs.toronto.edu/~kriz/cifar.html		
	– CIFAR-100	handwritten digits, each measuring 28 × 28 pixels – It encompasses color images of 10 distinct categories, each sized at 32 × 32 pixels – This dataset is a CIFAR-10 expansion, with 100 classes instead of the initial 10	– N/A	– N/A		
[77]	– CIFAR-10	– It was described by	– 60,000	– https://www.cs.toronto.edu/~kriz/cifar.html	√	Public
	– CIFAR-100	Perez Tobia et al. [76]	– N/A	– N/A		
	– TinyImageNet	– It was described by Perez Tobia et al. [76] – Image dataset	– N/A	– N/A		
[113]	– CK + – KDEF	– It is essentially a collection of staged face expressions	– 3,368 – 4,900	—	√	Private

(Continued)

Table 1: *Continued*

Ref.	Dataset name	Description	Size	Link of the dataset (if it is available)	Is the legally collected dataset?	Public/ Private
		– These are pictures of people's expressions on their faces				
[101]	– MNIST	– It was described by Perez Tobia et al. [76]	– 60,000	– -N/A	√	Public
[83]	– Fashion-MNIST		– 60,000	– https://github.com/zalandoresearch/fashion-mnist		
[107]	– CIFAR-10	– A collection of images. Every sample consists of a 28 by 28 grayscale picture labeled with one of 10 possible image classifications – It was described by Perez Tobia et al. [76]	– 60,000	https://www.cs.toronto.edu/~kriz/cifar.html		
[114]	– CIFAR-10 – MiniImageNet	– It was described by Perez Tobia et al. [76] – Image dataset	– 60,000 – N/A	– https://www.cs.toronto.edu/~kriz/cifar.html – N/A	√	Public
[78]	– CIFAR-10 – ImageNet-small	– It was described by Perez Tobia et al. [76] – Image dataset	– 60,000 – N/A	– https://www.cs.toronto.edu/~kriz/cifar.html – N/A	√	Public
[82]	– GTSRB	– Image dataset with 43 classes for traffic signs – It was described by Perez Tobia et al. [76]	– 51,830	https://github.com/lsw3130104597/Backdoor_detection	√	Public
[8]	– CIFAR-10		– 60,000	– N/A		
[79]	– Fed-EMNIST – Fashion MNIST – CIFAR-10	– The MNIST dataset was expanded in 2017 with the release of the federated extended modified NIST (Fed-EMNIST) dataset	– 280,000 – 60,000 – 60,000	– https://www.nist.gov/node/1298471/emnist-dataset – https://github.com/zalandoresearch/fashion-mnist https://www.cs.toronto.edu/~kriz/cifar.html	√	Public

(Continued)

Table 1: *Continued*

Ref.	Dataset name	Description	Size	Link of the dataset (if it is available)	Is the legally collected dataset?	Public/ Private
		– It was described by Kopcan et al. [101] – It was described by Perez Tobia et al. [76]				
[104]	– TEP – SP	– The Tennessee-Eastman process dataset is used in the creation, analysis, and assessment of industrial processes – The seven fault types in stainless steel that are not the typical kind are classified using the SP dataset	– 21 fault categories and one standard operating condition, with approximately 500 samples for each – 1941	– https://doi.org/10.1016/j.eng.2021.07.033	√	Public
[80]	– CIFAR-10	– It was described by Perez Tobia et al. [76]	– 60,000	– https://www.cs.toronto.edu/~kriz/cifar.html	√	Public
[81]	– ImageNe	– Image dataset	– 1,331,167	– N/A		
[94]	– eICU Collaborative Research Database – MNIST	– The MIT Laboratory for Computational Physiology and Philips Healthcare collaborated to create this medical dataset – It was described by Perez Tobia et al. [76]	– 30,000 – 60,000	– http://dx.doi.org/10.1038/sdata.2018.178 – N/A	√	Public
[105]	– MSCOCO – Flickr30K	– Image dataset – For NLP-related tasks, it is a frequently utilized large-scale dataset	– 123,287 – 31,000	– N/A	√	Public
[106]	– ISIC			– https://www.isic-archive.com	√	Public

(Continued)

Table 1: *Continued*

Ref.	Dataset name	Description	Size	Link of the dataset (if it is available)	Is the legally collected dataset?	Public/ Private
	– Messidor – ChestX-ray14	– The Dermoscopic Image dataset from the International Skin Imaging Collaboration (ISIC) for the categorization of melanoma4 – The Messidor dataset, includes numerical pictures of the posterior pole's eye fundus color – An X-ray picture collection called ChestX-ray14		– http://www.adcis.net/en/third-party/messidor – https://www.kaggle.com/c/ccc-chestx-ray14-multi-label-classication/data		
[52] [99]	CICIDS-2017	The Canadian Institute for Cybersecurity created this dataset, which is used for testing. There are 79 characteristics in all in this labeled dataset	170,360	– https://www.unb.ca/cic/datasets/ids-2017.html	√	Public
[97]	Drebin	It was described by Katebi et al. [103]				
[98]	ERMDS-X	The ERMDS-X dataset is a useful resource for assessing LB-MDS's resilience and making it easier to find any possible weak points in the system. There are 30,455 benign and 86,685 harmful samples in all	117,140	– https://github.com/lcjia94/	√	Public

(Continued)

Table 1: *Continued*

Ref.	Dataset name	Description	Size	Link of the dataset (if it is available)	Is the legally collected dataset?	Public/ Private
[84]	– MNIST – CIFAR-10	It was described by Perez Tobia et al. [76]				
[85]	– CIFAR-10 – CIFAR-100	It was described above in the study Perez Tobia et al. [76]				
[86]	– ImageNet-R	The ImageNet-R dataset, which contains over 15,000 reliable pictures, aims to support more study on the fascinating phenomena of image strength under assault. For objective benchmarking of adversarial attack and defensive techniques, this dataset is useful	15,000+	– https://github.com/elcronos/Defense-Friendly	√	Public
[87]	– FLICKR-25K – NUS-WIDE – MS-COCO	– FLICKR-25K is a collection of 25,000 Flickr photos with 38 labels added – NUS-WIDE comprises 81 concepts and 269,648 samples – It was described by Zhao et al. [105]	– 25,000 – 269,648 – 123,287	– https://doi.org/10.1145/1460096.1460104 – https://doi.org/10.1145/1646396.1646452 – https://doi.org/10.1007/978-3-319-10602-1_48	√	Public
[115]	– TEP – WM-811K	– It was described by Zhuo et al. [104] – The wafer map failure pattern identification is studied using an industrial wafer map dataset	– 21 fault categories and one standard operating condition, with approximately 500 samples for each – 811,457	– http://dx.doi.org/10.1016/0098-1354(93)80018-I – http://dx.doi.org/10.1109/TSM.2014.2364237	√	Private
[88]	– PaviaU		– 115 spectral bands	– http://www.ehu.eus/ccwintco/index.php/Hyperspectral	√	Public

(Continued)

Table 1: *Continued*

Ref.	Dataset name	Description	Size	Link of the dataset (if it is available)	Is the legally collected dataset?	Public/ Private
	– HoustonU 2018 – Salinas	– The ROSIS-03 satellite captured the Pavia University dataset covering the University of Pavia area, featuring a spatial size of 610 × 304, 115 spectral bands – The Houston University 2018 dataset2, acquired by CASI 1500 on Feb. 16, 2017, covers the University of Houston region, with a spatial size of 610 × 2,384, 48 spectral bands – The Salinas dataset, captured by AVIRIS over Salinas Valley, CA, USA, has a spatial size of 512 × 217, 204 spectral bands	– 48 spectral bands – 204 spectral bands	– https://hyperspectral.ee.uh.edu/?page – http://www.ehu.eus/ccwintco/index.php/Hyperspectral		
[89]	– MNIST – Fashion-MNIST	It was described by Kopcan et al. [101]				
[90]	– Cora – Citeseer – Pubmed	These datasets are utilized to conduct comparison studies between the suggested model and baselines made up of GCNs that are resilient to attacks			√	Private

(Continued)

Table 1: *Continued*

Ref.	Dataset name	Description	Size	Link of the dataset (if it is available)	Is the legally collected dataset?	Public/ Private
		and GCNs that are already in existence but did not take assaults into account. Contains: – Cora: 2,485 nodes, 10,138 edges, 7 classes, and 1,433 attribute – Citeseer: 2,110 nodes, 7,446 edges, 6 classes, and 3,703 attribute – Pubmed: 19,717 nodes, 88,651 edges, 5 classes, and 500 attribute				
[91]	—	There are 20,000 grayscale photos in all in this collection	20,000	—	√	Public
[92]	– Digits FEMNIST – CelebA	– Digits FEMINIST: The Federated Version of EMNIST Digits Dataset – CelebA is a dataset for image classification made up of photos of well-known faces each annotated with 40 binary features		– https://www.nist.gov/itl/products-and-services/emnist-dataset – http://mmlab.ie.cuhk.edu.hk/projects/CelebA.html	√	Public
[68]	– IMDB – AG's – SNLI	– The IMDB sentiment analysis dataset – News text classification dataset		—	√	Private

(Continued)

Table 1: *Continued*

Ref.	Dataset name	Description	Size	Link of the dataset (if it is available)	Is the legally collected dataset?	Public/ Private
		– Standard Natural Language Inference (SNLI)				
[100]	– Dataset-I – Dataset-II	– A total of eight fit and well volunteers (referred to as S1–S8, aged between 24 and 35) were enlisted to gather the data; three of the participants were female and five of the volunteers were male – Further testing of the detection framework was conducted using the preprocessed dataset DB-a of CapgMyo	– 2,400 – 1,600		√	Public
[69]	– IMDB – SST-2	– IMDB is a large movie review dataset that contains 25,000 training and 25,000 test samples – “Stanford’s sentiment tree collection,” SST-2, comprises 1,821 test samples, 872 verification samples, and 6,920 training samples	– 50,000 – 9,613		√	Private
[93]	– CHUK03		– 732		√	Public

(Continued)

Table 1: *Continued*

Ref.	Dataset name	Description	Size	Link of the dataset (if it is available)	Is the legally collected dataset?	Public/ Private
	– VIPER	– The 632 photos in the VIPER dataset, which is renowned for its difficulty in person reidentification, were taken by two nonoverlapping cameras – Six nonoverlapping cameras recorded 13,164 photos, or the IDs of 1,360 people, for the publicly available CUHK03 dataset	– 13,164			
[70]	– MNIST – Cifar10 – Cifar100 – FaceScrub – CelebA	– MNIST, Cifar10, and Cifar100 were described by Perez Tobia et al. [76] – This dataset contains over 100k face images of 530 individuals – CelebA was described by Rodríguez-Barroso et al. [92]	– FaceScrub: 106,863 images	– FaceScrub: vintage - resources (winklerbros.net)	√	Public
[71]	– MNIST – Cifar10 – Fashion-MNIST – SVHN	– MNIST, Cifar10, and Fashion-MNIST were described by Kopcan et al. [101] – SVHN: Extracted from house numbers detected from Google	– SVHN: 600,000 images	—	√	Public

(Continued)

Table 1: *Continued*

Ref.	Dataset name	Description	Size	Link of the dataset (if it is available)	Is the legally collected dataset?	Public/ Private
		Street View, with 32 × 32-pixel images				
[72]	– ImageNet – Imagenette – Places365	– ImageNet: is a large image database which has a great deal of images belonging to various classes and is suitable for the testing of DL – Imagenette: which is also a subset of ImageNet is comparatively smaller in size thereby making it easier to test the experiments quickly – Places365: dataset is a large-scale database of images that was developed to be used in training and testing ofScene Recognition and similar tasks in the field of computer vision	– 3.2 million images – N/A – 10 Million Image	– https://ieeexplore.ieee.org/abstract/document/5206848 – GitHub - fastai/imagenette: A smaller subset of 10 easily classified classes from Imagenet, and a little more French – Places: A 10 Million Image Database for Scene Recognition | IEEE Journals & Magazine | IEEE Xplore	√	- Private-Public-Private
[73]	– MNIST – CIFAR-10	It was described by Perez Tobia et al. [76]				
[109]	– Edge-IIoTset	The Edge-IIIoTset dataset is the consolidated data of all the layers of the tested, which include IoT and IIoT, for example, Cloud	– N/A	https://www.kaggle.com/datasets/mohamedamineferrag/edgeiiotset-cyber-security-dataset-of-iot-iiot	√	Public

(Continued)

Table 1: *Continued*

Ref.	Dataset name	Description	Size	Link of the dataset (if it is available)	Is the legally collected dataset?	Public/ Private
		computing, fog computing, block chain networks, and edges computing				
[95]	– X-IIoTID	Intersystem and connection independent dataset which eliminate the variation in data collected from different connectivity protocols, connected devices, and flow produced by the IIIoT network and system	– N/A	https://ieeexplore.ieee.org/abstract/document/9504604	√	Private
[102]		The dataset employed in this work contains malware and benign files. The malware are divided into five families: Locker, Mediyes, Winwebsec, Zbot, and Zeroaccess	– 8,970 malware and 3,140 benign files	https://figshare.com/articles/dataset/Malware_Detection_PEBased_Analysis_Using_Deep_Learning_Algorithm_Dataset/6635642	√	Public
[74]	– CelebA – CelebAHQ – FFHQ – ASIAWebFace – LFW	– CelebA dataset was described by Rodríguez-Barroso et al. [92] – CelebA-HQ dataset is as a subset of CelebA dataset which also offers Definition (HD) face images – FFHQ (Flickr-Faces-HQ) is high quality face dataset which contains High	– CelebAHQ: 30,000 (HD) face images – FFHQ: 70,000 HD face images – ASIAWebFace: about half a million face images of 10,575 people – LFW: 3,233 face images	—	√	Private

(Continued)

Table 1: *Continued*

Ref.	Dataset name	Description	Size	Link of the dataset (if it is available)	Is the legally collected dataset?	Public/ Private
		Definition (HD) face images at 1,024 × 1,024 resolution – CASIA-WebFace is based on the internet images and contains face images of 10,575 persons – Standard face study dataset is LFW (Labeled Faces in the Wild), which takes face images from real life scenes				
[111]	– CelebA	CelebA dataset was described by Rodríguez-Barroso et al. [92]				
[112]	– CIFAR10	– It was described by Perez Tobia et al. [76]				
	– ImageNet	– It was described by Yin et al. [72]				
[75]	– MNIST	It was described by Li et al. [70] and Al-Andoli et al. [71]				
	– Cifar10					
	– Cifar100					
	– SVHN					

Table 2: ML and DL techniques contribute to the defense against adversarial attacks

Ref.	ML methods	Optimization	Metrices
[122]	CNN	Optimizer = Adam, batch size = 64, and epoch = 250	Accuracy = 90%
[78]	DNN	Gradient descent over θ	Accuracy = 93%
[82]	ResNet18	Epoch = 10, Batch Size = 10, and Learning rate = 0.3	Accuracy = 67%
[79]	FL	Multi Krum with d = 20	Accuracy = 0.9652
[80]	Perceptual image ResNet-20	Optimizer = Adam, epoch = 20, learning rate = 0.001, and batch size = 128	Detected query rate = 95.6%
[94]	FL	Optimizing loss function	Accuracy = 94.759
[81]	DNN	Epoch = 50, learning rate = 0.2	Accuracy = 98.64
[83]	SVM	Kernel parameter based on Gaussian	Changes with different values of σ
[84]	DNN	Adding perturbations to input instances	FGSM = 97.30, PGD-10 = 96.63, and C&W = 97.13
[85]	DNN	Epoch = 200 and batch size	Accuracy = 86.52
[86]	CNN, DNN	The number of iterations needed	Accuracy, Precision, and Recall
[87]	SAAT	Reliable deep hashing models	t-MAP = 89.07
[88]	DNN	Activation function (ReLU)	Defense accuracy = 0.9924
[89]	OQFL	Convolution Layer, ReLU, and Batch normalization	Accuracy = 96.67
[90]	GCN	Activation function = SoftMax	Mean accuracy = 87%
[92]	FL	—	Accuracy = 0.9670
[68]	DNN	Batch Size	Accuracy = 92.82
[69]	DNN	—	Accuracy = 90%
[93]	CNN	—	Accuracy = 79.6%
[96]	Malware Detection Framework (MDF)	Optimizer = Adam, learning rates = 0.01, 0.001, 0.0001, 0.004, and epochs = 100, 150,200	Accuracy = 92.9
[123]	RNN, DNN	Using swarm optimization-based Artificial bee colony (ABC) algorithm	Accuracy, Precision, Recall, False Alarm, and *F*1-Score
[101]	AAE, DCGAN	Optimizers = Adam, epochs = 50, and batch size = 256	Error detection = 0.08% (AAE) and 1.89% (DCGAN)
[97]	DNN	The MalEAttack is optimized to achieve a high conversion rate (fooling rate)	Accuracy from 86.01% to 49.11% and fooling rate of 96.87%
[98]	MalConv and EMBER	Used obfuscation methods in generating samples that cause a decrease in the accuracy	accuracy reduction = 20%
[124]	ANN	3 hidden layers with the ReLU activation function, 51, 51 and 25 neurons, respectively and the ADAM optimizer. Batch size of 100 and just 10 epochs.	Precision, Recall, and *F*1-score
[99]	DT, XGB, LR, SVM and DNN	Activation functions = ReLU, SoftMax, and Dropout 0.01	*F*1 score = 0.93% and accuracy = 99.9%.
[100]	CNN	Activation functions = ReLU, SoftMax	Accuracy = 94.65
[8]	DNN	The optimal λ-value to decrease the information loss	Accuracy = 98.7 and 98.2%
[104]	DNN, KNN, and SVM	—	Accuracy = 75.8 and Confidence = 85.6
[105]	CNN	—	precision = 94.9, recall = 95.7
[106]	DNN	Optimizing on learn the distribution of perturbation with generative structure parameters.	Accuracy = 86.5% and AUC = 0.807
[107]	DNN	Reducing the number of examples needed to insert the backdoor	91.6% accuracy
[103]	Malware Clustering Approach	*Optimizer* = SGD, *Neurons* = Number of features, *Layers* = 5, *Error* = Back propagation and Activation = ReLU	Accuracy = 98.74, FPR = 8.68, and *F*1-score = 99.32
[110]	CNN	Optimizer = Adam, learning rate = 0.01 and weight decay 5e-4	Accuracy = 74.12
[113]	CNN	Epoch = 200 and learning rate = 0.01	Accuracy = 79%

(Continued)

Table 2: *Continued*

Ref.	ML methods	Optimization	Metrices
[114]	DnCNN	Handles the overfitting on the denoiser quite well and gives significant improvements on certified accuracy.	Standard Certified accuracy = 57.60 and Robust Certified accuracy = (27.20, 9.20, 2.20)
[115]	Spectral Normalized Gradient Descent (SNGD)	The network parameters θ are optimized	Accuracy = 94.39
[116]	DNN	Optimizer = SGD, epochs = 50, and learning rate = 0.01	Accuracy = 91.32
[71]	CNN, DNN, Fuzzy ARTMAP, Random Forest, K-Nearest Neighbors, XGBoost, or Gradient Boosting Machine	Optimize the performance of the DNN by minimizing the cross-entropy loss (LCE)	Accuracy, Precision, Recall, *F*1-score
[112]	DNN	Optimize the weight random switch strategies accordingly	Accuracy = 90.65%
[72]	DNN	Mask size to 64, the mask rate to 0.85, and configuring the Gaussianblur with a kernel size of 25 × 25 and a standard deviation of 11 × 11	Accuracy = 72%
[108]	LLMs	Non	Non
[109]	RF Classifier, CNN, and LSTM	Optimizer = Adam, Learning Rates: 0.001 for DL models and 0.01, Batch Sizes: 64 for DL models and 256 for traditional ML models, and Epochs = 50 epochs for DL models and 100 epochs for traditional ML models.	Accuracy = 0.85, 0.91, 0.93, Precision = 0.87, 0.92, 0.94, Recall = 0.82, 0.88, 0.89, *F*1-score = 0.84, 0.90, 0.91, and ROC AUC = 0.91, 0.94, 0.97
[75]	GAN	Activation function = ReLU, Sigmoid and BatchNorm2d = 16	Precision = 97.88, Recall = 97.87, and *F*1-score = 97.86
[102]	DNN	Optimizes the adversarial patch	Precision = 98.72, Recall = 98.87, and *F*1-score = 98.79
[73]	GAN	Optimizer = SGD, Learningrate = 0.001, Batch size = 32, = Trainingrounds = 30, and Aggregationalgorithm = FedProx	Accuracy = 93.07 Precision = 90.17, Recall = 94.10, and *F*1-score = 92.09
[74]	GAN	Epochs = 100, optimizer = Adam, and learning rate = 0.001	SSIM = 0.85, PSNR = 25.09, FID = 3.45, CSIM = 0.31, FDR = 99.19, and APR = 96.67
[95]	DAE	Optimizer = SGD, learning rate = 0.01, ReLU = activation function, and batch size of 32	Accuracy = 83.8 and *F*1-score = 62.1
[111]	GAN	Optimizer = Adam and epoch = 30	Precision = 0.948, recall = 0.938 = 0.943, and *F*1-score = 0.947
[70]	DNN	Optimizer = SGD, Adam, LR = 0.001,0.01 Batchsize = 128,64, and Epochs = 200,150	Accuracy = 96.32%
[125]	MLP, KNN, SVC, LR, NB, Meta Classifier, and DNN	N/A	Accuracy = 0.96, Precision = 0.96, Recall = 0.96, and *F*1-score = 0.96

amounts of high-quality data essential for training AI and ML algorithms. The process of collecting, labeling, and annotating data has proven to be both time-consuming and costly [118]. In addition, ethical considerations, potential biases, and the ramifications of AI-generated content demand meticulous evaluation, particularly in the domains of NLP, computer vision, and image analysis [119,120].

Conversely, ML involves several security risks that ML must contend with. For example, attackers have a strong motive to distort ML model outputs or obtain private information for their benefit [121]. To identify gaps in the application of ML techniques within the literature, we analyzed algorithms used in defense against adversarial attacks (refer to Table 2). This analysis enables us to pinpoint algorithms that have not been employed in prior studies, making them potential subjects for future research.

6.3 Future direction for protection in adversarial attack

The current landscape of protection against adversarial attacks also reveals several gaps and opens issues in related research. This need arises from the need to consider new evasion techniques that go beyond those in existence, evaluate not only the false positive but also the false negative rates of poisoning benign samples and finally develop the GAN implementations for both attack methods and protection solutions. The recommendations also include further RT, new attack, and protection strategy investigations, challenging the currently available models that cannot handle high-intensity attacks or complex target model taxa satisfactorily. Future works in different domains and concentrating on fields such as FL, model distillation systems, and GCNs call for improved robustness through a wide range of methodologies to address the challenges brought forth by data from disparate sources. Overall, research has continued to improve protection systems that are also trying to cover ever-changing adversarial attacks.

In relation to the current state in the field of healthcare, it is possible to identify a number of factors that may be involved when discussing the findings and developments of adversarial attacks and concerning triage services for patients with autism spectrum disorder (ASD) [126]. First, it concerns conducting research from available proofs to learn more about the challenges that can make the "ML" used for triaging ASD patients vulnerable to adversarial attacks. This entails assessing how these models mitigate adversarial influences to various extents and gaining insights into specific adversarial situations as well as potential adversary opportunities. Thus, as a gap, it is necessary to determine the steps for future work on how the selection of the parameters and features influences the likelihood of detecting certain adversarial examples of the ML models and act as indicators of the approaches that lead to highly robust models. Therefore, regardless of the current heuristic filtering, continuing to search for other protection mechanisms/countermeasures that, possibly, may be potentially relevant solely to ASD and related priorities for triage application could be useful and contribute to building a greater understanding of how it is possible to enhance the configuration of ML-based systems in terms of security and reliability. Therefore, more detailed studies should focus on enhancing awareness of the adversarial threats in relation to triage of patients with ASD and on pursuing the actions that can be taken to manage the possible adverse impacts, which at the same time will contribute to better functioning of the ML solution and its incorporation into healthcare organizations.

7 Conclusion

Adversarial attacks on ML require understanding the weaknesses of AI systems as part of a systematic investigation into protection strategies and practices against adversarial assaults. With the widespread use of ML and DL algorithms in almost every sphere, from finance and healthcare to AV cybersecurity, there is a potential impact that should not be discounted by adversarial attacks. The landscape of adversarial attacks on ML and DL models is changing very quickly, and there should be a wide range-based approach for countering them. This review provides an in-depth overview of the literature on the subject, detailing various avenues that attackers can take to compromise ML systems by exploiting vulnerabilities. Adversarial attacks include evasion and poisoning, both of which present major threats to the integrity and reliability of ML and DL applications. One recurring theme in the review is that a multiple-threat protection strategy is necessary. Alternatively, multiple strategies should be employed in parallel, from powerful architectures and training paradigms to include an anomaly detection system integrated into AT to increase the resistance of ML and DL models to potential attacks. Furthermore, this review provides insights into what researchers have been doing to increase the interpretability of ML and DL models. The most effective protection strategies should include the required components of accessibility and comprehension of model decisions so that stakeholders can identify weaknesses that could be used against them beforehand. With the continued development of adversarial ML, the associations among researchers, practitioners and policymakers have become essential. Advocating for standardized benchmarks, sharing datasets and committing to an open dialog will better enable the formation of more resilient and secure ML and DL systems. In addition, with the increasing societal

impact of ML and DL applications, ethical issues related to adversarial attacks and protection strategies must be carefully considered. This systematic review underscores the dynamic and rapidly changing nature of this conflict landscape in ML and DL. This emphasizes the urgent need to develop powerful protection plans that have a level of adaptability when facing new tactics adopted by enemies. We are in an age where ML and DL continue to transform the world in every possible way; hence, we have to focus on how to work toward building solid protection mechanisms that could assist us in maintaining the fidelity, safety, and robustness of these technologies under adversarial attacks.

Funding information: Authors state no funding involved.

Author contributions: Ghadeer Ghazi Shayea: writing–reviewing and editing. Mohd Hazli Mohammed Zabil: writing–reviewing and editing. Yahya Layth Khaleel: data curation, writing–original draft preparation, visualization, and supervision. Mustafa Abdulfattah Habeeb: visualization, investigation, and supervision. A.S. Albahri: conceptualization, methodology, supervision, and editing.

Conflict of interest: Authors state no conflict of interest.

Ethical approval: This article does not contain any studies with human participants or animals performed by any of the authors.

Informed consent to participate: Not applicable.

Consent to publication: Not applicable.

Data availability statement: Data sharing is not applicable to this article as no datasets were generated or analyzed during the current study. The study is based entirely on previously published data, which have been appropriately cited within the manuscript.

References

[1] Khaleel YL. Fake news detection using deep learning. Master Thesis. University of Miskolc, Hungary; 2021. p. 249–59. doi: 10.1007/978-3-030-91305-2_19.

[2] Mihna FKH, Habeeb MA, Khaleel YL, Ali YH, Al-Saeedi LAE. Using information technology for comprehensive analysis and prediction in forensic evidence. Mesop J Cybersecur. 2024;4(1):4–16. doi: 10.58496/MJCS/2024/002.

[3] Katarya R, Massoudi M. Recognizing fake news in social media with deep learning: a systematic review. 4th International Conference on Computer, Communication and Signal Processing, ICCCSP 2020. IEEE; 2020. p. 1–4. doi: 10.1109/ICCCSP49186.2020.9315255.

[4] Ongsulee P. Artificial intelligence, machine learning and deep learning. 2017 15th international conference on ICT and knowledge engineering (ICT&KE). IEEE; 2017. p. 1–6.

[5] Campesato O. Artificial intelligence, machine learning, and deep learning. Mercury Learning and Information; 2020.

[6] Habeeb MA. Hate speech detection using deep learning master thesis. Master Thesis. University of Miskolc, Hungary; 2021. http://midra.uni-miskolc.hu/document/40792/38399.pdf.

[7] Zhang L, Ghosh R, Dekhil M, Hsu M, Liu B. Combining lexicon-based and learning-based methods for twitter sentiment analysis. HP Lab Tech Rep. 2011;89(89):1–8. https://api.semanticscholar.org/CorpusID:16228540.

[8] Husnoo MA, Anwar A. Do not get fooled: Defense against the one-pixel attack to protect IoT-enabled Deep Learning systems. Ad Hoc Netw. 2021;122:102627. doi: 10.1016/j.adhoc.2021.102627.

[9] Habeeb MA, Khaleel YL, Albahri AS. Toward smart bicycle safety: leveraging machine learning models and optimal lighting solutions. In: Daimi K, Al Sadoon A, editors. Proceedings of the Third International Conference on Innovations in Computing Research (ICR'24). Cham: Springer Nature Switzerland; 2024. p. 120–31.

[10] Khaleel YL, Habeeb MA, Albahri AS, Al-Quraishi T, Albahri OS, Alamoodi AH. Network and cybersecurity applications of defense in adversarial attacks: A state-of-the-art using machine learning and deep learning methods. J Intell Syst. 2024;33(1):20240153. doi: 10.1515/jisys-2024-0153.

[11] Hasan Z, Mohammad HR, Jishkariani M. Machine learning and data mining methods for cyber security: a survey. Mesop J Cybersecur. 2022;2022:47–56. doi: 10.58496/MJCS/2022/006.
[12] Paya A, Arroni S, García-Díaz V, Gómez A. Apollon: A robust defense system against adversarial machine learning attacks in intrusion detection systems. Comput Secur. 2024;136:103546. doi: 10.1016/j.cose.2023.103546.
[13] Ali G, Mijwil MM, Buruga BA, Abotaleb M, Adamopoulos I. A survey on artificial intelligence in cybersecurity for smart agriculture: state-of-the-art, cyber threats, artificial intelligence applications, and ethical concerns. Mesop J Comput Sci. 2024;2024:71–121. doi: 10.58496/MJCSC/2024/007.
[14] Mustaffa SNFNB, Farhan M. Detection of false data injection attack using machine learning approach. Mesop J Cybersecur. 2022;2022:38–46. doi: 10.58496/MJCS/2022/005.
[15] Akhtar N, Mian A, Kardan N, Shah M. Advances in adversarial attacks and defenses in computer vision: a survey. IEEE Access. 2021;9:155161–96. doi: 10.1109/ACCESS.2021.3127960.
[16] Zhang WE, Sheng QZ, Alhazmi A, Li C. Adversarial attacks on deep-learning models in natural language processing. ACM Trans Intell Syst Technol. 2020;11(3):1–41. doi: 10.1145/3374217.
[17] Cai K, Zhu X, Hu Z. Reward poisoning attacks in deep reinforcement learning based on exploration strategies. Neurocomputing. 2023;553:126578. doi: 10.1016/j.neucom.2023.126578.
[18] Echeberria-Barrio X, Gil-Lerchundi A, Mendialdua I, Orduna-Urrutia R. Topological safeguard for evasion attack interpreting the neural networks' behavior. Pattern Recognit. 2024;147:110130. doi: 10.1016/j.patcog.2023.110130.
[19] Jodayree M, He W, Janicki R. Preventing image data poisoning attacks in federated machine learning by an encrypted verification key. Procedia Comput Sci. 2023;225:2723–32. doi: 10.1016/j.procs.2023.10.264.
[20] Macas M, Wu C, Fuertes W. Adversarial examples: A survey of attacks and defenses in deep learning-enabled cybersecurity systems. Expert Syst Appl. Mar. 2024;238:122223. doi: 10.1016/j.eswa.2023.122223.
[21] Devabhakthini P, Parida S, Shukla RM, Nayak SC. Analyzing the impact of adversarial examples on explainable machine learning. arXiv Prepr arXiv230708327. 2023.
[22] Fang J, Jiang Y, Jiang C, Jiang ZL, Liu C, Yiu SM. State-of-the-art optical-based physical adversarial attacks for deep learning computer vision systems. Expert Syst Appl. 2024;250:123761. doi: 10.1016/j.eswa.2024.123761.
[23] Kong Z, Xue J, Wang Y, Huang L, Niu Z, Li F. A survey on adversarial attack in the age of artificial intelligence. Wirel Commun Mob Comput. 2021;2021(1):4907754. doi: 10.1155/2021/4907754.
[24] Chen T, Ling J, Sun Y. White-box content camouflage attacks against deep learning. Comput Secur. 2022;117:102676. doi: 10.1016/j.cose.2022.102676.
[25] Eyas A, Engstrom L, Athalye A, Lin J. Black-box adversarial attacks with limited queries and information. In: Dy J, Krause A, editors. 35th International Conference on Machine Learning, ICML 2018. International Machine Learning Society (IMLS); 2018. p. 3392–401. https://proceedings.mlr.press/v80/ilyas18a.html.
[26] Apruzzese G, Subrahmanian VS. Mitigating adversarial gray-box attacks against phishing detectors. IEEE Trans Dependable Secur Comput. 2023;20(5):3753–69. doi: 10.1109/TDSC.2022.3210029.
[27] Wang Y, Wu Y, Wu S, Liu X, Zhou W, Zhu L, et al. Boosting the transferability of adversarial attacks with frequency-aware perturbation. IEEE Trans Inf Forensics Secur. 2024;19:6293–304. doi: 10.1109/TIFS.2024.3411921.
[28] Zhang Y, Ruan W, Wang F, Huang X. Generalizing universal adversarial perturbations for deep neural networks. Mach Learn. 2023;112(5):1597–626. doi: 10.1007/s10994-023-06306-z.
[29] Bostani H, Moonsamy V. EvadeDroid: A practical evasion attack on machine learning for black-box Android malware detection. Comput Secur. 2024;139:103676. doi: 10.1016/j.cose.2023.103676.
[30] Randhawa RH, Aslam N, Alauthman M, Khalid M, Rafiq H. Deep reinforcement learning based Evasion Generative Adversarial Network for botnet detection. Futur Gener Comput Syst. 2024;150:294–302. doi: 10.1016/j.future.2023.09.011.
[31] Shi L, Chen Z, Shi Y, Zhao G, Wei L, Tao Y, et al. Data poisoning attacks on federated learning by using adversarial samples. Proceedings - 2022 International Conference on Computer Engineering and Artificial Intelligence, ICCEAI 2022. 2022. p. 158–62. doi: 10.1109/ICCEAI55464.2022.00041.
[32] Tolpegin V, Truex S, Gursoy ME, Liu L. Data poisoning attacks against federated learning systems. In: Chen L, Li N, Liang K, Schneider S, editors. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). Cham: Springer International Publishing; 2020. p. 480–501. doi: 10.1007/978-3-030-58951-6_24.
[33] Chakraborty A, Alam M, Dey V, Chattopadhyay A, Mukhopadhyay D. A survey on adversarial attacks and defences. CAAI Trans Intell Technol. 2021;6(1):25–45. doi: 10.1049/cit2.12028.
[34] Wang T, Zhu L, Zhang Z, Zhang H, Han J. Targeted adversarial attack against deep cross-modal hashing retrieval. IEEE Trans Circuits Syst Video Technol. 2023;33(10):6159–72. doi: 10.1109/TCSVT.2023.3263054.
[35] Sagduyu YE, Erpek T, Ulukus S, Yener A. Is semantic communication secure? a tale of multi-domain adversarial attacks. IEEE Commun Mag. 2023;61(11):50–5. doi: 10.1109/MCOM.006.2200878.
[36] Aldahdooh A, Hamidouche W, Déforges O. Revisiting model's uncertainty and confidences for adversarial example detection. Appl Intell. 2023;53(1):509–31. doi: 10.1007/s10489-022-03373-y.
[37] Yan A, Huang T, Ke L, Liu X, Chen Q, Dong C. Explanation leaks: Explanation-guided model extraction attacks. Inf Sci. 2023;632:269–84. doi: 10.1016/j.ins.2023.03.020.
[38] Park C, Kim Y, Park JG, Hong D, Seo C. Evaluating differentially private generative adversarial networks over membership inference attack. IEEE Access. 2021;9:167412–25. doi: 10.1109/ACCESS.2021.3137278.

[39] Anthi E, Williams L, Javed A, Burnap P. Hardening machine learning denial of service (DoS) defences against adversarial attacks in IoT smart home networks. Comput Secur. 2021;108:102352. doi: 10.1016/j.cose.2021.102352.
[40] Seo S, Lee Y, Kang P. Cost-free adversarial defense: Distance-based optimization for model robustness without adversarial training. Comput Vis Image Underst. 2023;227:103599. doi: 10.1016/j.cviu.2022.103599.
[41] Guo C, Rana M, Cissé M, Van Der Maaten L. Countering adversarial images using input transformations. 6th International Conference on Learning Representations, ICLR 2018 - Conference Track Proceedings, International Conference on Learning Representations, ICLR. 2018. https://openreview.net/forum?id=SyJ7ClWCb.
[42] McDaniel, Wu X, Jha S, Swami A. Distillation as a defense to adversarial perturbations against deep neural networks. In Proceedings - 2016 IEEE Symposium on Security and Privacy, SP 2016. 2016. p. 582–97. doi: 10.1109/SP.2016.41.
[43] Hang J, Han K, Chen H, Li Y. Ensemble adversarial black-box attacks against deep learning systems. Pattern Recognit. 2020;101:107184. doi: 10.1016/j.patcog.2019.107184.
[44] Li D, Li Q. Adversarial deep ensemble: evasion attacks and defenses for malware detection. IEEE Trans Inf Forensics Secur. 2020;15:3886–900. doi: 10.1109/TIFS.2020.3003571.
[45] Zheng Y, Velipasalar S. Part-based feature squeezing to detect adversarial examples in person re-identification networks. In Proceedings - International Conference on Image Processing. ICIP; 2021. p. 844–8. doi: 10.1109/ICIP42928.2021.9506511.
[46] Shaham U, Yamada Y, Negahban S. Understanding adversarial training: Increasing local stability of supervised models through robust optimization. Neurocomputing. 2018;307:195–204. doi: 10.1016/j.neucom.2018.04.027.
[47] Schwartz D, Alparslan Y, Kim E. Regularization and Sparsity for Adversarial Robustness and Stable Attribution. In: Bebis G, Yin Z, Kim E, Bender J, Subr K, Kwon BC, et al., editors. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). Cham: Springer International Publishing; 2020. p. 3–14. doi: 10.1007/978-3-030-64556-4_1.
[48] Panda P, Roy K. Implicit adversarial data augmentation and robustness with Noise-based Learning. Neural Netw. 2021;141:120–32. doi: 10.1016/j.neunet.2021.04.008.
[49] Anand D, Tank D, Tibrewal H, Sethi A. Self-supervision vs. transfer learning: robust biomedical image analysis against adversarial attacks. In Proceedings - International Symposium on Biomedical Imaging. 2020. p. 1159–63. doi: 10.1109/ISBI45749.2020.9098369.
[50] Siva Kumar RS, Nystrom M, Lambert J, Marshall A, Goertzel M, Comissoneru A, et al. Adversarial machine learning-industry perspectives. In Proceedings - 2020 IEEE Symposium on Security and Privacy Workshops, SPW 2020. 2020. p. 69–75. doi: 10.1109/SPW50608.2020.00028.
[51] Li W, Tug S, Meng W, Wang Y. Designing collaborative blockchained signature-based intrusion detection in IoT environments. Futur Gener Comput Syst. 2019;96:481–9. doi: 10.1016/j.future.2019.02.064.
[52] Roshan K, Zafar A, Ul Haque SB. Untargeted white-box adversarial attack with heuristic defence methods in real-time deep learning based network intrusion detection system. Comput Commun. 2024;218:97–113. doi: 10.1016/j.comcom.2023.09.030.
[53] Siu K, Moitra A, Li M, Durling M, Herencia-Zapana H, Interrante J, et al. Architectural and behavioral analysis for cyber security. In AIAA/IEEE Digital Avionics Systems Conference – Proceedings. 2019. p. 1–10. doi: 10.1109/DASC43569.2019.9081652.
[54] Kuppa A, Grzonkowski S, Asghar MR, Le-Khac NA. Black box attacks on deep anomaly detectors. In ACM International Conference Proceeding Series, in ARES '19. New York, NY, USA: Association for Computing Machinery; 2019. doi: 10.1145/3339252.3339266.
[55] Apruzzese G, Colajanni M, Ferretti L, Marchetti M. Addressing adversarial attacks against security systems based on machine learning. In International Conference on Cyber Conflict. CYCON; 2019. p. 1–18. doi: 10.23919/CYCON.2019.8756865.
[56] DIng X, Fang H, Zhang Z, Choo KKR, Jin H. Privacy-preserving feature extraction via adversarial training. IEEE Trans Knowl Data Eng. 2022;34(4):1967–79. doi: 10.1109/TKDE.2020.2997604.
[57] Ali Z, Mohammed A, Ahmad I. Vulnerability of deep forest to adversarial attacks. IEEE Trans Inf Forensics Secur. 2024;19:5464–75. doi: 10.1109/TIFS.2024.3402309.
[58] Owezarski P. Investigating adversarial attacks against Random Forest-based network attack detection systems. In Proceedings of IEEE/IFIP Network Operations and Management Symposium 2023, NOMS 2023. 2023. p. 1–6. doi: 10.1109/NOMS56928.2023.10154328.
[59] Mustapha A, Khatoun R, Zeadally S, Chbib F, Fadlallah A, Fahs W, et al. Detecting DDoS attacks using adversarial neural network. Comput Secur. 2023;127:103117. doi: 10.1016/j.cose.2023.103117.
[60] Sohrabi C, Franchi T, Mathew G, Kerwan A, Nicola M, Griffin M, et al. PRISMA 2020 statement: What's new and the importance of reporting guidelines. Int J Surg. 2021;88:105918. Elsevier. doi: 10.1016/j.ijsu.2021.105918.
[61] Khaw KW, Alnoor A, Al-Abrrow H, Tiberius V, Ganesan Y, Atshan NA. Reactions towards organizational change: a systematic literature review. Curr Psychol. 2023;42:19137–60. doi: 10.1007/s12144-022-03070-6.
[62] Albahri AS, Duhaim AM, Fadhel MA, Alnoor A, Baqer NS, Alzubaidi L, et al. A systematic review of trustworthy and explainable artificial intelligence in healthcare: Assessment of quality, bias risk, and data fusion. Inf Fusion. 2023;96:156–91. doi: 10.1016/j.inffus.2023.03.008.
[63] Albahri AS, Khaleel YL, Habeeb MA, Ismael RD, Hameed QA, Deveci M, et al. A systematic review of trustworthy artificial intelligence applications in natural disasters. Comput Electr Eng. 2024;118:109409. doi: 10.1016/j.compeleceng.2024.109409.
[64] Albahri AS, Alwan JK, Taha ZK, Ismail SF, Hamid RA, Zaidan AA, et al. IoT-based telemedicine for disease prevention and health promotion: State-of-the-Art. J Netw Comput Appl. 2021;173:102873. doi: 10.1016/j.jnca.2020.102873.

[65] Rusydiana AS. Bibliometric analysis of journals, authors, and topics related to COVID-19 and Islamic finance listed in the Dimensions database by Biblioshiny. Sci Ed. 2021;8(1):72–8. doi: 10.6087/kcse.232.
[66] Jadeja M, Shah K. Tree-map: A visualization tool for large data. In CEUR Workshop Proceedings. 2015. p. 9–13. http://ceur-ws.org/Vol-1393/paper-07.pdf.
[67] Zaidan AA, Alnoor A, Albahri OS, Mohammed RT, Alamoodi AH, Albahri AS, et al. Review of artificial neural networks-contribution methods integrated with structural equation modeling and multi-criteria decision analysis for selection customization. Eng Appl Artif Intell. 2023;124:106643. doi: 10.1016/j.engappai.2023.106643.
[68] Shi J, Li L, Zeng D. ASCL: Adversarial supervised contrastive learning for defense against word substitution attacks. Neurocomputing. 2022;510:59–68. doi: 10.1016/j.neucom.2022.09.032.
[69] Shao K, Yang J, Ai Y, Liu H, Zhang Y. BDDR: An effective defense against textual backdoor attacks. Comput Secur. 2021;110:102433. doi: 10.1016/j.cose.2021.102433.
[70] Li P, Huang J, Wu H, Zhang Z, Qi C. SecureNet: Proactive intellectual property protection and model security defense for DNNs based on backdoor learning. Neural Netw. 2024;174:106199. doi: 10.1016/j.neunet.2024.106199.
[71] Al-Andoli MN, Tan SC, Sim KS, Goh Y, Lim CP. A framework for robust deep learning models against adversarial attacks based on a protection layer approach. IEEE Access. 2024;12:17522–40. doi: 10.1109/ACCESS.2024.3354699.
[72] Yin L, Wang S, Wang Z, Wang C, Zhan D. Attribution guided purification against adversarial patch. Displays. 2024;83:102720. doi: 10.1016/j.displa.2024.102720.
[73] Abdel-Basset M, Hawash H, Moustafa N, Razzak I, Abd Elfattah M. Privacy-preserved learning from non-i.i.d data in fog-assisted IoT: A federated learning approach. Digit Commun Netw. 2024;10(2):404–15. doi: 10.1016/j.dcan.2022.12.013.
[74] Zhang Y, Fang Y, Cao Y, Wu J. RBGAN: Realistic-generation and balanced-utility GAN for face de-identification. Image Vis Comput. 2024;141:104868. doi: 10.1016/j.imavis.2023.104868.
[75] Luo Y, Zhu T, Liu Z, Mao T, Chen Z, Pi H, et al. GANFAT: Robust federated adversarial learning with label distribution skew. Futur Gener Comput Syst. 2024;160:711–23. doi: 10.1016/j.future.2024.06.030.
[76] Perez Tobia J, Braun P, Narayan A. AGS: attribution guided sharpening as a defense against adversarial attacks. In Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), The University of British Columbia. Kelowna, Canada: Springer Science and Business Media Deutschland GmbH; 2022. p. 225–36. doi: 10.1007/978-3-031-01333-1_18.
[77] Hwang D, Lee E, Rhee W. AID-purifier: A light auxiliary network for boosting adversarial defense. Neurocomputing. 2023;541:126251. doi: 10.1016/j.neucom.2023.126251.
[78] Dai T, Feng Y, Chen B, Lu J, Xia ST. Deep image prior based defense against adversarial examples. Pattern Recognit. 2022;122:108249. doi: 10.1016/j.patcog.2021.108249.
[79] Rodríguez-Barroso N, Martínez-Cámara E, Luzón MV, Herrera F. Dynamic defense against byzantine poisoning attacks in federated learning. Futur Gener Comput Syst. 2022;133:1–9. doi: 10.1016/j.future.2022.03.003.
[80] Choi SH, Shin J, Choi YH. PIHA: Detection method using perceptual image hashing against query-based adversarial attacks. Futur Gener Comput Syst. 2023;145:563–77. doi: 10.1016/j.future.2023.04.005.
[81] Li Y, Wu B, Feng Y, Fan Y, Jiang Y, Li Z, et al. Semi-supervised robust training with generalized perturbed neighborhood. Pattern Recognit. 2022;124:108472. doi: 10.1016/j.patcog.2021.108472.
[82] Lu S, Li R, Liu W, Chen X. Defense against backdoor attack in federated learning. Comput Secur. 2022;121:102819. doi: 10.1016/j.cose.2022.102819.
[83] Li W, Liu X, Yan A, Yang J. Kernel-based adversarial attacks and defenses on support vector classification. Digit Commun Netw. 2022;8(4):492–7. doi: 10.1016/j.dcan.2021.12.003.
[84] Hassanin M, Radwan I, Moustafa N, Tahtali M, Kumar N. Mitigating the impact of adversarial attacks in very deep networks. Appl Soft Comput. 2021;105:107231. doi: 10.1016/j.asoc.2021.107231.
[85] Liu J, Jin Y. Multi-objective search of robust neural architectures against multiple types of adversarial attacks. Neurocomputing. 2021;453:73–84. doi: 10.1016/j.neucom.2021.04.111.
[86] Pestana C, Liu W, Glance D, Mian A. Defense-friendly images in adversarial attacks: Dataset and metrics for perturbation difficulty. Proceedings - 2021 IEEE Winter Conference on Applications of Computer Vision, WACV 2021, IEEE Winter Conference on Applications of Computer Vision (WACV) CL-Electr Network. 35 Stirling Hwy, Crawley, WA 6009, Australia: Univ Western Australia; 2021. p. 556–65. doi: 10.1109/WACV48630.2021.00060.
[87] Yuan X, Zhang Z, Wang X, Wu L. Semantic-aware adversarial training for reliable deep hashing retrieval. IEEE Trans Inf Forensics Secur. 2023;18:4681–94. doi: 10.1109/TIFS.2023.3297791.
[88] Shi C, Liu Y, Zhao M, Pun CM, Miao Q. Attack-invariant attention feature for adversarial defense in hyperspectral image classification. Pattern Recognit. 2024;145:109955. doi: 10.1016/j.patcog.2023.109955.
[89] Yamany W, Moustafa N, Turnbull B. OQFL: An optimized quantum-based federated learning framework for defending against adversarial attacks in intelligent transportation systems. IEEE Trans Intell Transp Syst. 2023;24(1):893–903. doi: 10.1109/TITS.2021.3130906.
[90] Lee Y, Han SW. CAGCN: Causal attention graph convolutional network against adversarial attacks. Neurocomputing. 2023;538:126187. doi: 10.1016/j.neucom.2023.03.048.
[91] Zha H, Zhang W, Yu N, Fan Z. Enhancing image steganography via adversarial optimization of the stego distribution. Signal Process. 2023;212:109155. doi: 10.1016/j.sigpro.2023.109155.

[92] Rodríguez-Barroso N, Martínez-Cámara E, Luzón MV, Herrera F. Backdoor attacks-resilient aggregation based on Robust Filtering of Outliers in federated learning for image classification. Knowl Syst. 2022;245:108588. doi: 10.1016/j.knosys.2022.108588.
[93] Kanwal S, Shah JH, Khan MA, Nisa M, Kadry S, Sharif M, et al. Person re-identification using adversarial haze attack and defense: A deep learning framework. Comput Electr Eng. 2021;96:107542. doi: 10.1016/j.compeleceng.2021.107542.
[94] Nair AK, Sahoo J, Raj ED. Privacy preserving Federated Learning framework for IoMT based big data analysis using edge computing. Comput Stand Interfaces. 2023;86:103720. doi: 10.1016/j.csi.2023.103720.
[95] Gungor O, Rosing T, Aksanli B. "ROLDEF: RObust layered defense for intrusion detection against adversarial attacks. In Proceedings -Design, Automation and Test in Europe, DATE. Institute of Electrical and Electronics Engineers Inc; 2024. https://doi.org/10.23919/date58400.2024.10546886.
[96] Shaukat K, Luo S, Varadharajan V. A novel method for improving the robustness of deep learning-based malware detectors against adversarial attacks. Eng Appl Artif Intell. 2022;116:105461. doi: 10.1016/j.engappai.2022.105461.
[97] Rathore H, Nandanwar A, Sahay SK, Sewak M. Adversarial superiority in android malware detection: Lessons from reinforcement learning based evasion attacks and defenses. Forensic Sci Int Digit Investig. 2023;44:301511. doi: 10.1016/j.fsidi.2023.301511.
[98] Jia L, Yang Y, Tang B, Jiang Z. ERMDS: A obfuscation dataset for evaluating robustness of learning-based malware detection system. BenchCouncil Trans Benchmarks, Stand Eval. 2023;3(1):100106. doi: 10.1016/j.tbench.2023.100106.
[99] Lin YD, Pratama JH, Sudyana D, Lai YC, Hwang RH, Lin PC, et al. ELAT: Ensemble learning with adversarial training in defending against evaded intrusions. J Inf Secur Appl. 2022;71:103348. doi: 10.1016/j.jisa.2022.103348.
[100] Xue B, Wu L, Liu A, Zhang X, Chen X, Chen X. Detecting the universal adversarial perturbations on high-density sEMG signals. Comput Biol Med. 2022;149:105978. doi: 10.1016/j.compbiomed.2022.105978.
[101] Kopcan J, Škvarek O, Klimo M. Anomaly detection using autoencoders and deep convolution generative adversarial networks. Transp Res Procedia. 2021;55:1296–303. doi: 10.1016/j.trpro.2021.07.113.
[102] Zhan D, Duan Y, Hu Y, Li W, Guo S, Pan Z. MalPatch: evading DNN-based malware detection with adversarial patches. IEEE Trans Inf Forensics Secur. 2024;19:1183–98. doi: 10.1109/TIFS.2023.3333567.
[103] Katebi M, Rezakhani A, Joudaki S. ADCAS: adversarial deep clustering of android streams. Comput Electr Eng. 2021;95:107443. doi: 10.1016/j.compeleceng.2021.107443.
[104] Zhuo Y, Shardt YAW, Ge Z. One-variable attack on the industrial fault classification system and its defense. Engineering. 2022;19:240–51. doi: 10.1016/j.eng.2021.07.033.
[105] Zhao M, Wang B, Guo W, Wang W. Protecting by attacking: A personal information protecting method with cross-modal adversarial examples. Neurocomputing. 2023;551:126481. doi: 10.1016/j.neucom.2023.126481.
[106] Xu M, Zhang T, Li Z, Liu M, Zhang D. Towards evaluating the robustness of deep diagnostic models by adversarial attack. Med Image Anal. 2021;69:101977. doi: 10.1016/j.media.2021.101977.
[107] Soremekun E, Udeshi S, Chattopadhyay S. Towards backdoor attacks and defense in robust machine learning models. Comput Secur. 2023;127:103101. doi: 10.1016/j.cose.2023.103101.
[108] Charfeddine M, Kammoun HM, Hamdaoui B, Guizani M. ChatGPT's security risks and benefits: offensive and defensive use-cases, mitigation measures, and future implications. IEEE Access. 2024;12:30263–310. doi: 10.1109/ACCESS.2024.3367792.
[109] Khaleel TA. Developing robust machine learning models to defend against adversarial attacks in the field of cybersecurity. In HORA 2024 - 6th International Congress on Human-Computer Interaction, Optimization and Robotic Applications, Proceedings. Institute of Electrical and Electronics Engineers Inc; 2024. doi: 10.1109/HORA61326.2024.10550799.
[110] Meng L, Jiang X, Wu D. Adversarial robustness benchmark for EEG-based brain–computer interfaces. Futur Gener Comput Syst. 2023;143:231–47. doi: 10.1016/j.future.2023.01.028.
[111] Ul Ghani MAN, She K, Rauf MA, Alajmi M, Ghadi YY, Algarni A. Securing synthetic faces: A GAN-blockchain approach to privacy-enhanced facial recognition. J King Saud Univ - Comput Inf Sci. 2024;36(4):102036. doi: 10.1016/j.jksuci.2024.102036.
[112] Wei X, Wang X, Yan Y, Jiang N, Yue H. ALERT: A lightweight defense mechanism for enhancing DNN robustness against T-BFA. J Syst Archit. 2024;152:103160. doi: 10.1016/j.sysarc.2024.103160.
[113] Shehu HA, Browne WN, Eisenbarth H. An anti-attack method for emotion categorization from images[Formula presented]. Appl Soft Comput. 2022;128:109456. doi: 10.1016/j.asoc.2022.109456.
[114] Nayak GK, Rawal R, Chakraborty A. DE-CROP: Data-efficient Certified Robustness for Pretrained Classifiers. In Proceedings - 2023 IEEE Winter Conference on Applications of Computer Vision, WACV 2023, Indian Institute of Science, Department of Computational and Data Sciences. Bangalore, India: Institute of Electrical and Electronics Engineers Inc.; 2023. p. 4611–20. doi: 10.1109/WACV56688.2023.00460.
[115] Yin Z, Zhuo Y, Ge Z. Transfer adversarial attacks across industrial intelligent systems. Reliab Eng Syst Saf. 2023;237:109299. doi: 10.1016/j.ress.2023.109299.
[116] Cheng Z, Zhu F, Zhang XY, Liu CL. Adversarial training with distribution normalization and margin balance. Pattern Recognit. 2023;136:109182. doi: 10.1016/j.patcog.2022.109182.
[117] Dadvandipour S, Khaleel YL. Application of deep learning algorithms detecting fake and correct textual or verbal news. Prod Syst Inf Eng. 2022;10(2):37–51. doi: 10.32968/psaie.2022.2.4.
[118] Hassan A, Mahmood A. Efficient deep learning model for text classification based on recurrent and convolutional layers. In Proceedings - 16th IEEE International Conference on Machine Learning and Applications, ICMLA. 2017. p. 1108–13. doi: 10.1109/ICMLA.2017.00009.

[119] Albahri AS, Khaleel YL, Habeeb MA. The considerations of trustworthy AI components in generative AI; A Letter to Editor. Appl Data Sci Anal. 2023;2023:108–9. doi: 10.58496/adsa/2023/009.
[120] Ray PP. ChatGPT: A comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope. Internet Things Cyber-PhysSyst. 2023;3:121–54. doi: 10.1016/j. iotcps. 2023.04. 003.
[121] Wang C, Chen J, Yang Y, Ma X, Liu J. Poisoning attacks and countermeasures in intelligent networks: Status quo and prospects. Digit Commun Netw. 2022;8(2):225–34. doi: 10.1016/j.dcan.2021.07.009.
[122] Hou X, Liu J, Xu B, Wang X, Liu B, Qiu G. Class-aware domain adaptation for improving adversarial robustness. Image Vis Comput. 2020;99:103926. doi: 10.1016/j.imavis.2020.103926.
[123] Qureshi AUH, Larijani H, Mtetwa N, Yousefi M, Javed A. An adversarial attack detection paradigm with swarm optimization. In Proceedings of the International Joint Conference on Neural Networks, glasgow caledonian university, school of computing, Engineering and Built Environment. Glasgow, United Kingdom: Institute of Electrical and Electronics Engineers Inc.; 2020. doi: 10.1109/IJCNN48605.2020.9207627.
[124] Pawlicki M, Choraś M, Kozik R. Defending network intrusion detection systems against adversarial evasion attacks. Futur Gener Comput Syst. 2020;110:148–54. doi: 10.1016/j.future.2020.04.013.
[125] Ojo S, Krichen M, Alamro MA, Mihoub A. TXAI-ADV: Trustworthy XAI for Defending AI models against adversarial attacks in realistic CIoT. Electron. 2024;13(9):1769. doi: 10.3390/electronics13091769.
[126] Shayea GG, Zabil M, Albahri AS, Joudar SS, Hamid RA, Albahri OS, et al. Fuzzy evaluation and benchmarking framework for robust machine learning model in real-time autism triage applications. Int J Comput Intell Syst. 2024;17(1):151. doi: 10.1007/s44196-024-00543-3.