Applied Data Science and Analysis
Vol.2024, **pp**. 121–147
DOI: https://doi.org/10.58496/ADSA/2024/011; ISSN: 3005-317X
https://mc04.manuscriptcentral.com/adsa

Review Article

Adversarial Attacks in Machine Learning: Key Insights and Defense Approaches

Yahya Layth Khaleel [1], *, , Mustafa Abdulfattah Habeeb[1], , Hussein Alnabulsi [2],

[1] *College of Computer Science and Mathematics, Tikrit University, Iraq*

[2] *School of Information Technology and Engineering, Melbourne Institute of Technology, Melbourne, 3001, Australia*

ARTICLEINFO

Article History

Received 18 May 2024
Revised 02 Jul 2024
Accepted 19 Jul 2024
Published 07 Aug 2024

Keywords

Adversarial Robustness

Adversarial Defense

Security

AI Ethics

Threat Detection

ABSTRACT

There is a considerable threat present in genres such as machine learning due to adversarial attacks which include purposely feeding the system with data that will alter the decision region. These attacks are committed to presenting different data to machine learning models in a way that the model would be wrong in its classification or prediction. The field of study is still relatively young and has to develop strong bodies of scientific research that would eliminate the gaps in the current knowledge. This paper provides the literature review of adversarial attacks and defenses based on the highly cited articles and conference published in the Scopus database. Through the classification and assessment of 128 systematic articles: 80 original papers and 48 review papers till May 15, 2024, this study categorizes and reviews the literature from different domains, such as Graph Neural Networks, Deep Learning Models for IoT Systems, and others. The review posits findings on identified metrics, citation analysis, and contributions from these studies while suggesting the area's further research and development for adversarial robustness' and protection mechanisms. The identified objective of this work is to present the basic background of adversarial attacks and defenses, and the need for maintaining the adaptability of machine learning platforms. In this context, the objective is to contribute to building efficient and sustainable protection mechanisms for AI applications in various industries

1. INTRODUCTION

AI stands for artificial intelligence and it is considered as a pinnacle of the modern software development accomplishments in the sphere of computer science [1]. This domain has evolved over time with several aspects and application hence a large number of scholarly articles have been developed to analyze the prospects of this domain [2]–[5]. Among the countless subdivisions belonging to AI are NLP [6], [7], computer vision [8].

A critical branch of AI is machine learning (ML), which' refers to the creation of algorithms and statistical models that can be used to teach computers specific tasks without being told exactly what to do [9]. These systems acquire experience and are dynamic in solving difficult classification patterns and most of the data pattern and tasks [10]–[12]. Machine learning can be seen as the main building block for many AI use cases and continued to push forward developments in predictive models, recommendation engines, and more [13].

Nevertheless, as the integration of AI systems and, in particular, the application of ML methods intensified, new problems cropped up, one of which is adversarial attacks [14]. Adversarial attacks are basically acts of endeavoring to feed wrong information to the AI models with an aim of making them arrive at the wrong decision or even wrong classification. They operate on the themes of the machine learning algorithms and as a result cause considerable weakening of the performance in activities such as image recognition, speech and route identification.

The analysis of adversarial attacks is important because it demonstrates the flaws of AI systems and encourages researchers to work on the designs that are less vulnerable. In this way, the researcher would know how adversarial inputs can lead to the degradation of the ML results. as well as its corresponding reliability, hence moving closer to the development of higher levels of secure and trustworthy AI [15]. The example of AI's interaction with security shows that the field is constantly developing, and that sustained research is critical since new risks continue to emerge, as will be further discussed, for AI to be deployed safely and effectively.

**Corresponding author. Email: yahya@tu.edu.iq*

In this study, we reviewed of the literature including articles and conference proceeding that present related information and research findings. To methodology, we selected papers published in the Scopus database that got over 100 citations so as to only include the most impactful literature. This led to the identification of 80 published original research papers and 48 review papers up to May 15, 2024. The classification of these papers was carried out by a relatively simple one-layer protocol focusing on the identification of major results and conclusions. Our findings are summarized with the help of several figures showing the papers' classification, the total number of citations, and the average number of citations per topic.

As seen in the reviewed literature, topics within the adversarial attacks and defense domain are very broad. Among them is the study of Graph Neural Networks (GNNs) [16] that deals with weaknesses of neural network data for graphs and the strategies for defending against adversarial manipulations. Another appreciable branch comprises Deep Learning Models[17] , with subsequent research studies looking into adversarial attacks effects on healthcare systems and face recognition systems, and encoding of radio signals, and automatic speech recognition systems.

Despite the growing body of research on adversarial attacks, several gaps remain unaddressed. One significant gap is the lack of comprehensive understanding of the mechanisms underlying adversarial vulnerabilities in different types of AI models. While there has been substantial progress in identifying and categorizing different types of attacks, there is still limited knowledge about why certain models are more susceptible than others and how these vulnerabilities can be systematically mitigated. Additionally, much of the existing research focuses on specific domains, such as image recognition, leaving other critical areas like natural language processing and graph-based learning less explored. Another gap lies in the effectiveness and generalizability of current defense mechanisms. Many proposed defenses are often tailored to specific types of attacks or models and may not provide robust protection across various scenarios. Addressing these gaps is crucial to developing more resilient AI systems.

The significance of this study represents due to the Adversarial Attacks profound impact on the reliability and security of AI systems. As AI becomes increasingly integrated into various aspects of society, from healthcare and finance to autonomous driving and national security, the potential consequences of adversarial attacks grow correspondingly severe. This not only undermines the trust in AI systems but also poses significant risks in real-world applications where errors can lead to catastrophic outcomes, such as misdiagnoses in healthcare, financial fraud, or accidents involving autonomous vehicles. Understanding and mitigating these attacks is therefore essential to ensure the safe and effective deployment of AI technologies.

Altogether, this review is expected to leaving no stone unturned in presenting a panoramic view of all the existing work done in the adversarial attack and defense domain of Machine learning. To sum up, in this paper, we discussed key research findings in the area of AI security, analyzed areas for further research, and outlined the possible development of the directions in the future.

Thus, the purpose of this study is to map the high cited research in adversarial attacks and defenses against AI with a focus on overarching research gaps. In addition, the following specific objectives relates to the general objective of the study:

Analyze the Impact of Adversarial Attacks: The adversarial attacks threaten many of the AI applications which will demonstrate their effects on the stabilities and securities of the system.

Review and Classify Research: Conduct a meta-analysis of the literature and categorize them according to the research topics, the approaches and outcomes achieved.

Identify Research Gaps: Stress what does not achieve or explain about the state of the art in everything that can make an opponent vulnerable or the weak points of the existing protection measures.

Suggest Future Research Directions: Propose the possible ways of the further study that can contribute to the creation of more effective and generally applicable defense strategies.

Therefore, the attainment of these objectives will assist the study to participate in improvements of AI security, in providing information and results helpful to researchers and practitioner involved in developing AI security against adversarial attacks.

2. AN OVERVIEW OF ADVERSARIAL ATTACKS

Machine learning models make precise predictions based on the learned statistical representations of their input data [18]. Models that form the cornerstone in areas such as image and audio classification, machine translation, optical character recognition, automated driving, and question answering have impressive predictive capabilities [19]. Often, these models are robust to small perturbations in their input. Adversarial examples are inputs to machine learning models that are designed to intentionally cause the model to make a mistake [20]. The creation of adversarial examples and the perturbed data itself are examples of adversarial attacks. In situations such as medical diagnoses, facial authentication systems, and autonomous vehicles, even a small percentage of adversarial examples causing models to incorrectly predict can be harmful [21], [22].

The perturbations are created via heuristic optimization to satisfy two conditions: small perturbation and misclassification. Adversarial attacks can happen at various stages throughout the lifetime of a model, in training via poisoned data, of transferability property, during the black box where the model has to be queried, and in real-time once the model has been deployed via test-time evasion [23], [24]. Detecting and defending against such attacks is an emerging area of research, and solutions from previous works should be expanded upon or refined. However, the generation of adversarial examples for defense against adversarial attacks is the motivation for our work on this robust and reproducible adversarial example generator [25-26].

Although there exist several forms of adversarial examples, adversarial examples can be predominantly classified into the following categories based on different aspects of their attributes. Classifying these attacks is based on the knowledge that the white box or black box attacks possess. This implies that an adversary knows the system completely or partially. According to the types of knowledge about the model, attacks can be divided into three categories.

i) White Box Attacks: In this type of attack, the adversary has complete access and knowledge of the underlying model. It knows all the model parameters and its architectures [27].

ii) Black Box Attacks: In such type of attack, the adversary is agnostic about the underlying model, i.e., it has no knowledge about its architecture, parameters, or even the training data. Rather, it only possesses the access to the output of the given input data [28].

iii) Transfer-based Attacks: In such attacks, an adversary has access to a substitute model even if the underlying model is unknown [29].

Most researchers categorize their methods as being white-box (aware of the model parameters and their values) or black-box (unaware of the model parameters and their values) [30], [31]. Given an adversarial perturbation, it is important to determine whether it may be imperceptible to a human, affect the performance of the machine learning model, or cause errors during transfer between models or domains [32]. Furthermore, evaluation metrics take into account defense mechanisms and compare both models and methods of attack.

The success rate of adversarial attacks quantifies the percentage of attacks caused by adversarial perturbations, while a low success rate indicates a failure to fool a particular model. Distorted transfers describe a transferable adversarial attack's success, with lower distortion rates resulting in greater transferable adversarial attacks [33]. An approach can be considered difficult to assess by low transferability. Although accuracy is a common evaluation tool across a variety of methodologies, it fails in adversarial context. Given the rarity of distortions within a dataset, the accuracy of adversarial inputs is vulnerable to many attacks [34], [35]. A high-perturbation attack with low accuracy, however, may be considered more powerful than a reduced-perturbation attack with high accuracy. When two methods disrupt perturbations or tend to find the weakest elements of the model, our robustness evaluation is contrasted with them [36].

3. ADVERSARIAL ATTACK: AN ANALYSIS AND TAXONOMY

An adversarial attack is specified through the contrast of deliberate aims to cheat or falsify in machine learning via the injection of deliberately created testing data by an expert, sometimes different from the actual input [37]. Adversarial attacks can be explained as an intent of undermining the dataset that can alter the decision borders of machine’s response. The redresses of these attacks can be diverse and, for instance, they can map on the data to special inputs which could become a triggering mechanism for misclassification or wrong predictions [27], [38-41] .

Adversarial attacks research is probably the most important area, which also scientists together with other practitioners have to pay more attention to. The role, however, is not minimal, but the field is still at its early state. This is why the technique of revealing the knowledge gap and understanding the field in general has to be done with great care.

This entailed an analysis of the operational environment that distinguished it from others and combined the factors that most research cite. Analyzing the philosophical allows one to perceive the main trends and create a solid basis for the next stage of study. Our goal was to achieve this. So, we carefully explored all scholarly articles and conference proceedings in which adversarial attacks were discussed in detail. However, we discarded those publications that do not have more than 100 citations from the Scopus database.

The outcomes of our search go with 80 of original research papers from the beginning to the 15 May 2024 in Scopus website. We finished up with a much-simplified version of the paper’s descriptions and, therefore, classified them accordingly. Using straightforward classification protocol based on one layer, we then searched the content of the documents, pointing to the main highlights and conclusions. These findings were displayed in Figures 1-3 subsequently.

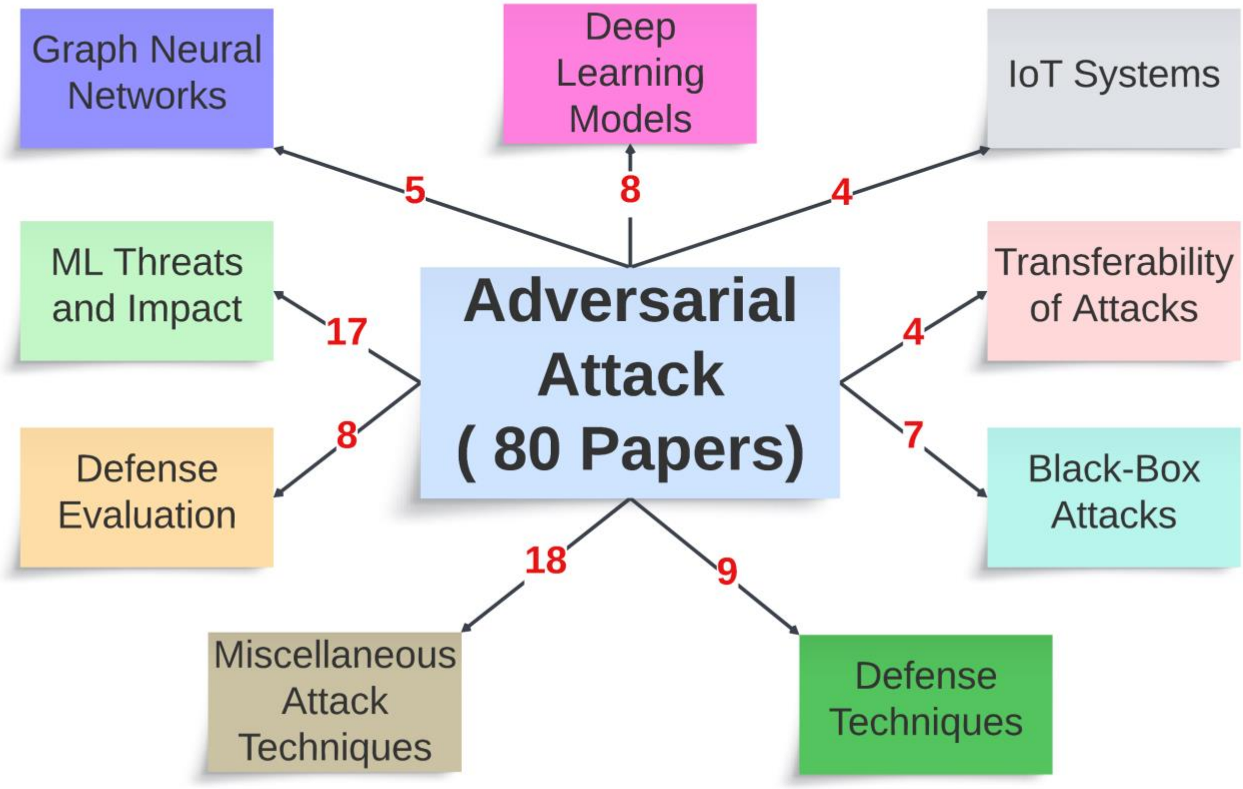

Fig. 1. The taxonomy of original papers in adversarial attack that have more than 100 citations

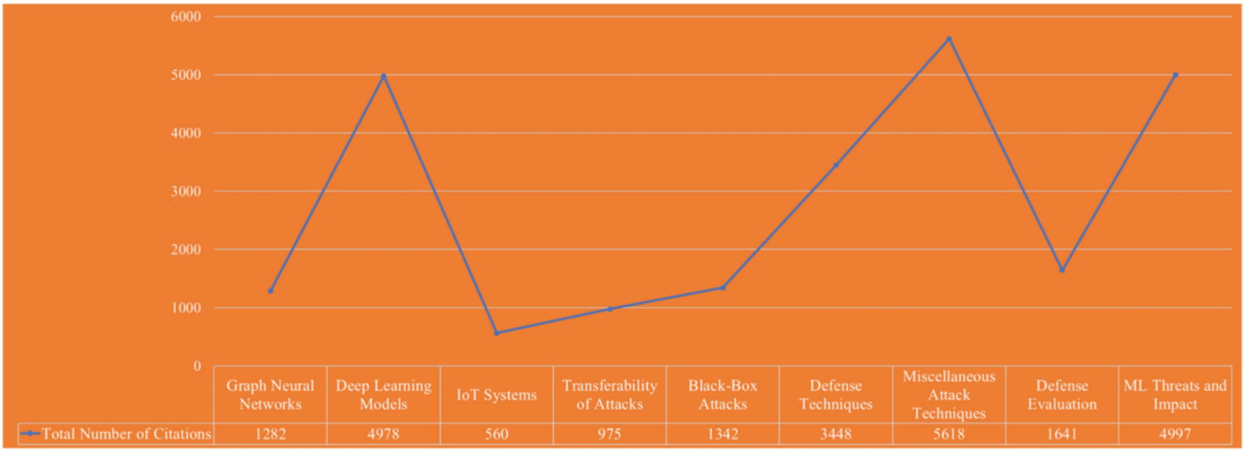

Fig. 2. Total Number of citations for each topic

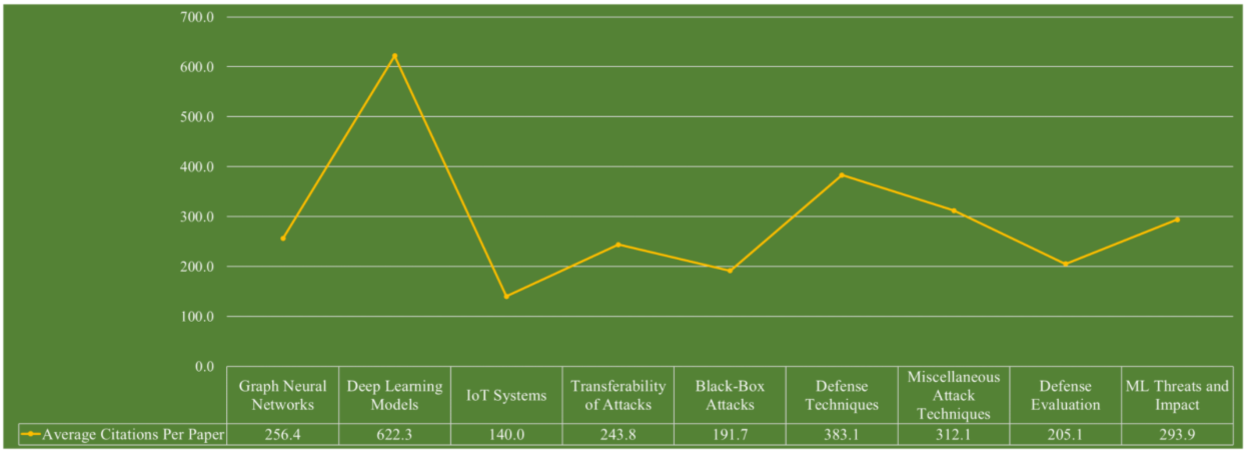

Fig. 3. Average Number of citations per paper for each topic

This study presents a comprehensive and inclusive analysis, which is comprised of all research papers, adversarial attacks and protection from different fields and their appropriate number of citations. The materials illustrate broadly the research diversity and the academical world's level of impact shown by other studies.

3.1 Graph Neural Networks

Research in adversarial attacks and the methods used in Graph Neural Networks (GNNs) is really labor-intensive. The most critical aspect of this research is is general vulnerabilities in the neural network data for graph, which in fact the how much attention can be paid on this area. Whereas the other research, by investigating the same attribute, look into the vulnerability of GNN to adversarial perturbations and suggest techniques to counter the problem. The second type of work that is very much essential also deals with offensive mechanisms and tries to focus more on attack mechanisms on graphs, explicating the theoretical issues as well as providing the real practice in order to defend the GNNs. Figure 4 provide more details about the number of citations in this topic [16], [42-45] .

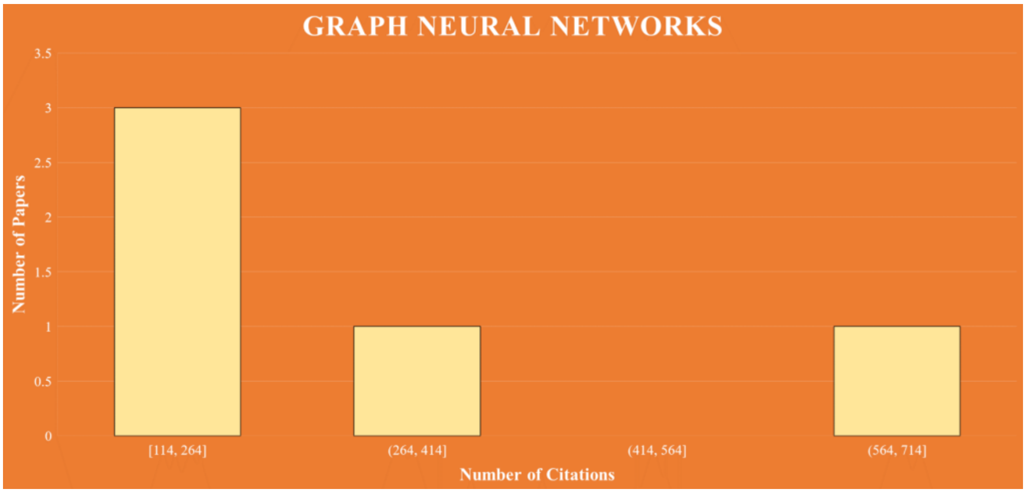

Fig. 4. Number of citations in Graph Neural Networks.

3.2 Deep Learning Models

The section of the best papers entails journals identified with significant changes that boost model protection against adversarial attacks, among others. Along with the general potential applications mentioned above, the specific practical implication of malicious attacks for healthcare contexts, as well as the real face recognition systems cases, are discussed below in more detail. Investigation samples the ways of both radio signal categories and automatic speech recognition systems as an adversarial attack method, suggesting the higher flexibility of the adversarial research studies in deep learning. Oppositional risk management is an activity of defending a wide range of deep learning issues that are embedded within the survey. Figure 5 provide more details about the number of citations in this topic [46-53].

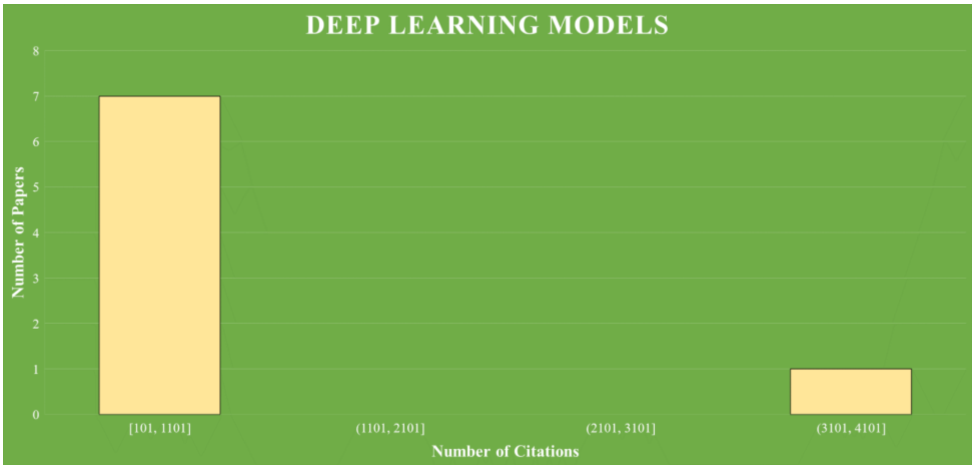

Fig. 5. Number of citations in Deep Learning Models.

3.3 IoT Systems

In the IoT environment the nature of adversarial attacks is unique and complex. Many papers are published that do research in this security area and outline the detection of intrusion and to secure the networks, which are eased by systems of network intrusion detection. This leads to lower citation counts, probably for research that has been focused on security of vulnerable networks thereby exposing IoT platforms to attackers. These studies are the strong pillars on which the security systems of IoT stand for the they cope with the security holes on a vast scale. Figure 6 provide more details about the number of citations in this topic [17], [54-56].

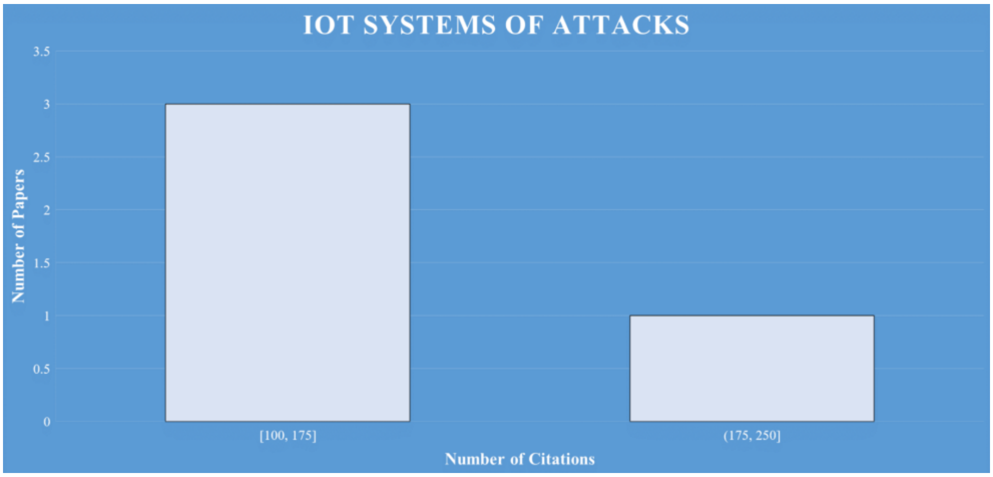

Fig. 6. Number of citations in IoT Systems.

3.4 Transferability of Attacks

Model transferability is a specific concern with an adversarial attack that has less protecting a model against attack customized for another model. Some of the meaningful papers devote to a detailed analysis of the attack transferability and the methods that improve the diffusion of the attacks giving additional insights to theoretical mechanisms of this process. Transfer attacks between diverse models would underline the weakness of protection, thus, it is crucial to understand why this happens as to find more comprehensive protection in future. This subfield is going to the core of the transferability issue, trying to pinpoint the factors that enable it and develop techniques to tackle the associated risks. Figure 7 provide more details about the number of citations in this topic [57- 60].

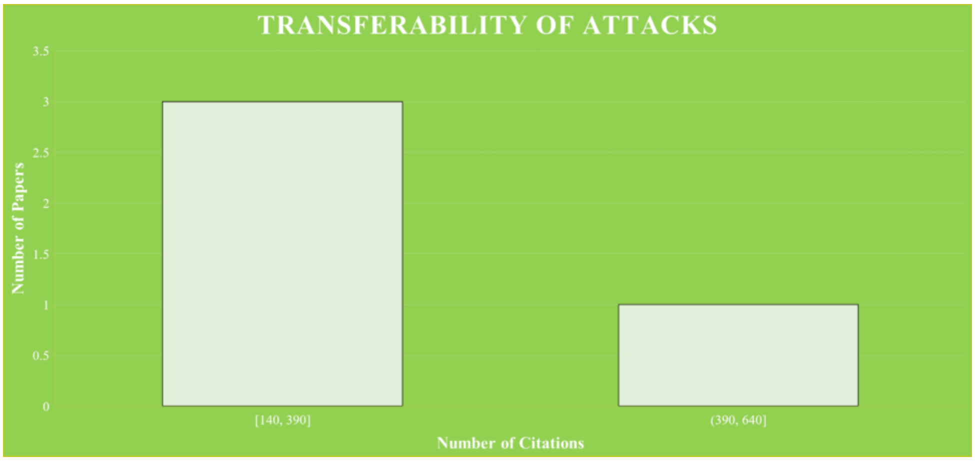

Fig. 7. Number of citations in Transferability of Attacks.

3.5 Black-Box Attacks

Black hypotheses, where the attack does not require that the assailant is in possession of the target model in question, pose a fundamental and an unsettling problem on which the researchers must work. Studies showing distinct ways to manipulate automatic surveillance techniques, various anti computing strategies and other methods to scout the system is a part of noteworthy findings. These papers provide us with precisely the sort of real-life implications that black-box attacks have and what kind of reliable protection mechanisms we should be pursuing. This field of study is interested in creating solutions that can prevent adversarial attacks even in the best case when the information about the attack is limited. Figure 8 provide more details about the number of citations in this topic [61-67].

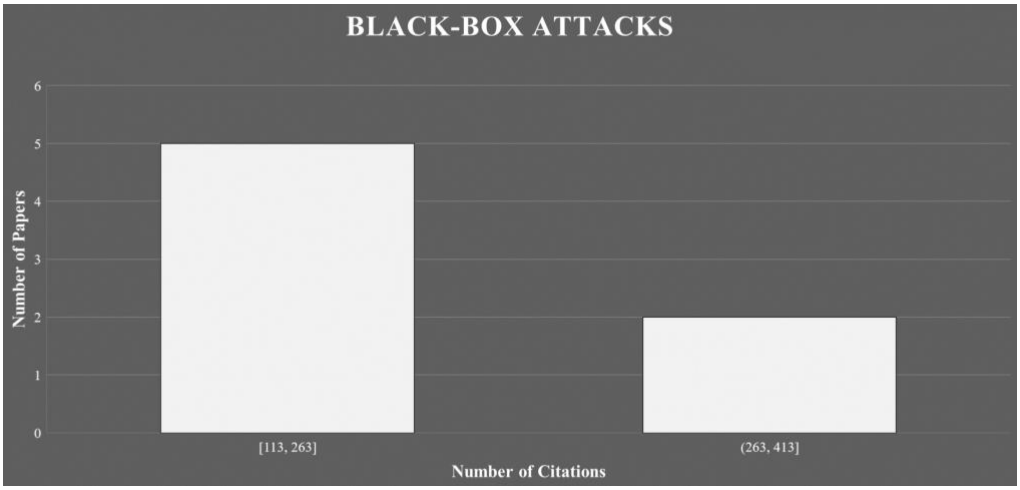

Fig. 8. Number of citations in Black-Box Attacks.

3.6 Defense Techniques

The highly quoted section deals with the broad range of papers that solely responsible for the strength of varied defense strategies against the adversarial attacks. The research focuses on ensemble training approaches, trustworthy determines of the robustness, and the use of generative models for the resistant. Other research teams undertake feature dispersal-based adversarial training and have integrated defenses provided by automatic malware detection systems. These measures enable better machine learning systems' assault protection against attacks that are becoming more advanced. Figure 9 provide more details about the number of citations in this topic [68–76].

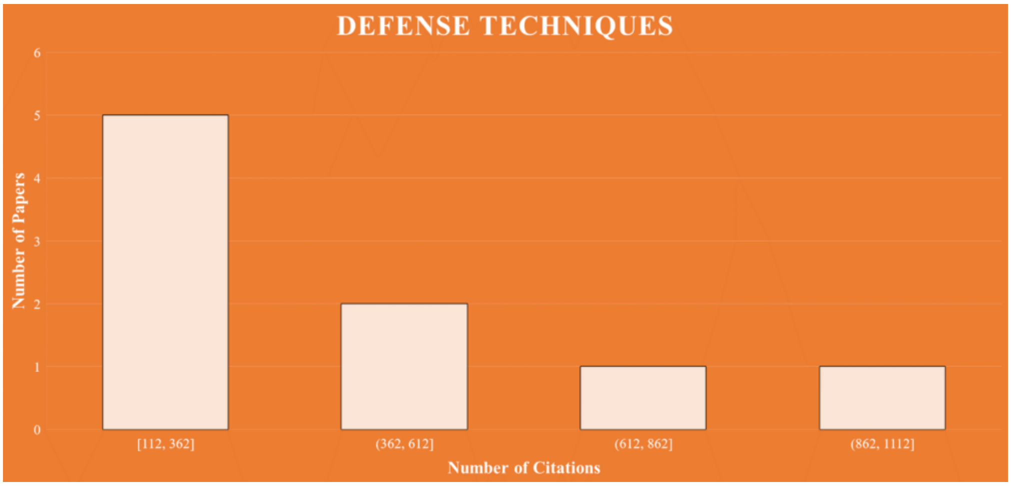

Fig. 9. Number of citations in Defense Techniques.

3.7 Miscellaneous Attack Techniques

In this section, we learn about a range of attack approaches as well as cases of unpronounceable syllables and assaults on pattern recognition systems. Here we subdivide the area into research on an assault of sensors by self-driving cars, as well as developing better transferability of adversarial attacks and a technology of adversarial camouflage. Equally, these investigations show that there are numerous ways from which the attackers can take advantage, suggesting a need for elaborated protection require deploying varied applications and typical conditions. Figure 10 provide more details about the number of citations in this topic [21], [77–93].

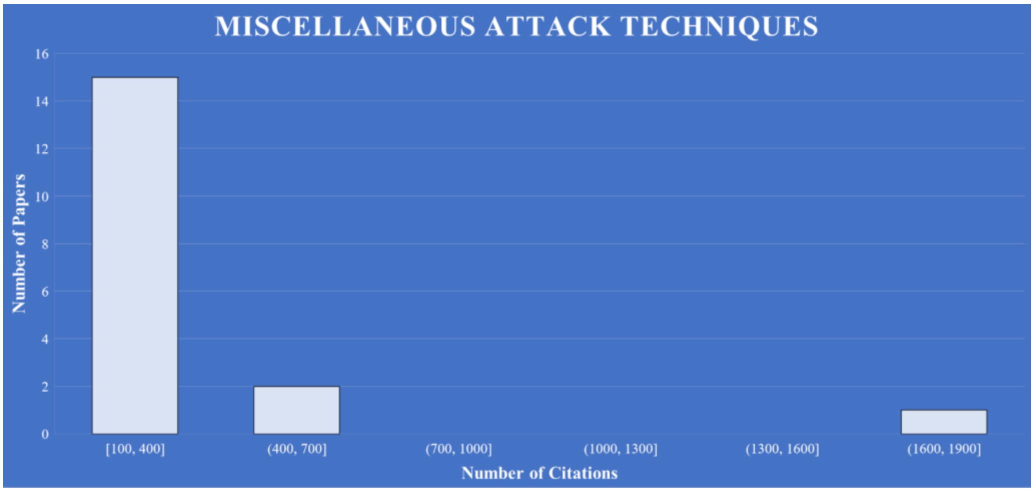

Fig. 10. Number of citations in Miscellaneous Attack Techniques.

3.8 Defense Evaluation

This block investigates the appropriate defense mechanism's sturdiness against adversarial assault. Significant articles are adapted to study the effectiveness of attacks on adaptive systems, the observability of linear systems under attack, and general reviews of defense methodologies for deep learning. In addition, the research studies enhancing strong graph convolutional network structure as well as defenses against the black-box membership inference attacks are also made. Assessing defense mechanisms is of utmost importance in order to appreciate their strengths and weaknesses, furthered corrections concerning adversarial resilience. Figure 11 provide more details about the number of citations in this topic [39], [94–100].

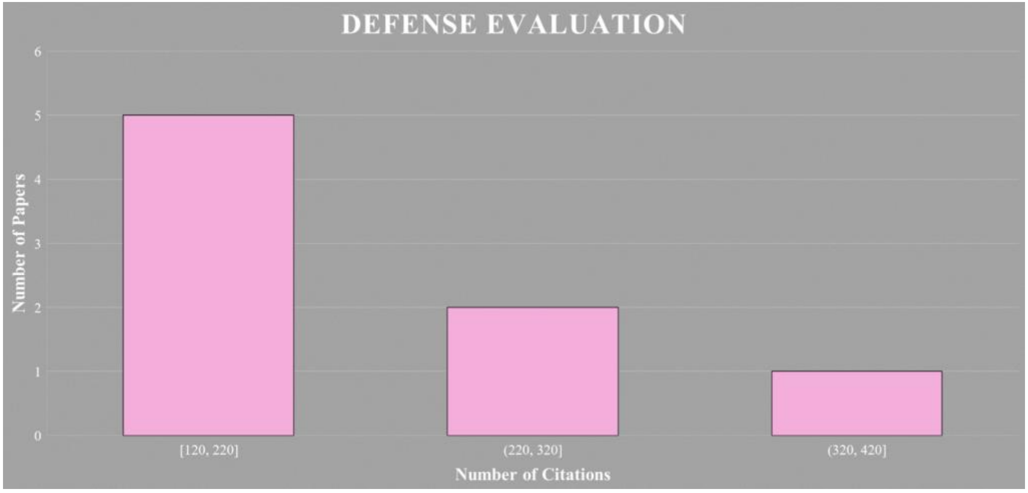

Fig. 11. Number of citations in Defense Evaluation.

3.9 ML Threats and Impact

In this part, scholars question the systemic risks and potential dangers of the misbehavior of machine learning systems. The central group focuses on adversarial risks, the dangers in judging on weak attacks, and the approaches to achieve robust perception under vehicles driving. The research also embodies robust estimation and control of cyber-physical systems and secure physical adversarial attacks. In doing this, the works shed the light on the widespread consequences of adversarial attacks and how it is essential to design strong and reliable attack preventive systems capable of repelling various types of adversaries. Figure 12 provide more details about the number of citations in this topic [101–117].

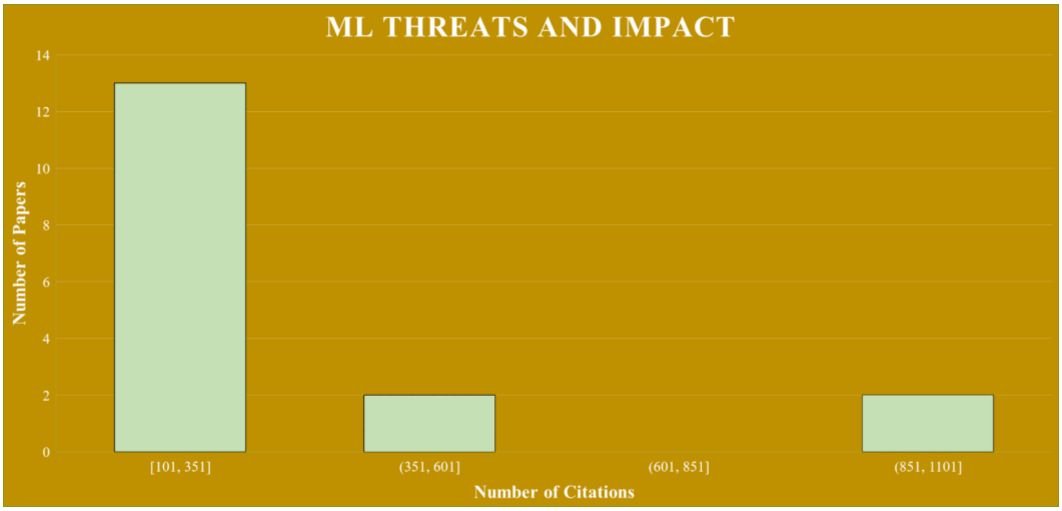

Fig. 12. Number of citations in ML Threats and Impact.

In a nutshell, coming from a vast domain of research---adversarial attacks and protection ---which is yet to have the answer on crucial gaps and needs to be addressed as soon as possible by scholars. On the other hand, the theoretical base and methods including protection machine learning as well as simulation exercises are comprehensively covered but it becomes a subject for further research in practice of the implementation and actual deployment of such defenses. Also, investigations of the role of adverse robustness in many other areas (such as, IoT systems, industrial control systems, etc.) are just getting started. These are just the common entry points to quite sophisticated attacks currently exploiting those same systems which have become so important and widely deployed. While adversarial attack transferability has been already well documented in theory, to this day, it remains unclear in a practical sense how different machine learning models and the scenarios observed in the real world are affected by this phenomenon. However, black-box attacks and protection are not yet investigated in detail, and therefore, systems that use proprietary and closed-source applications decrease the level of occurrence of such cases in commercial activities. Moreover, it undoubtedly gives rise to many national security issues that remain either not dealt with them at all or not sufficient, affecting health care and on-board vehicles to name a few, so which requires multidisciplinary work to look not only on technology but also ethics and regulations. Evolving problems arise, that is, a permanently changing prophylaxis is a requirement. However, the directional research can only be seen as reflected in some literature, hence the more efforts in this field are needed for the emergence of adaptive and resilient machine learning systems. This will see models deployed with no chances of gaps and the safety be maintained in real-time.

On the other hand, the same Scopus search was performed for all the reviews and surveys that were relevant. Finally, the total number was 48 papers which involve other taxonomy as shows in Figure 13.

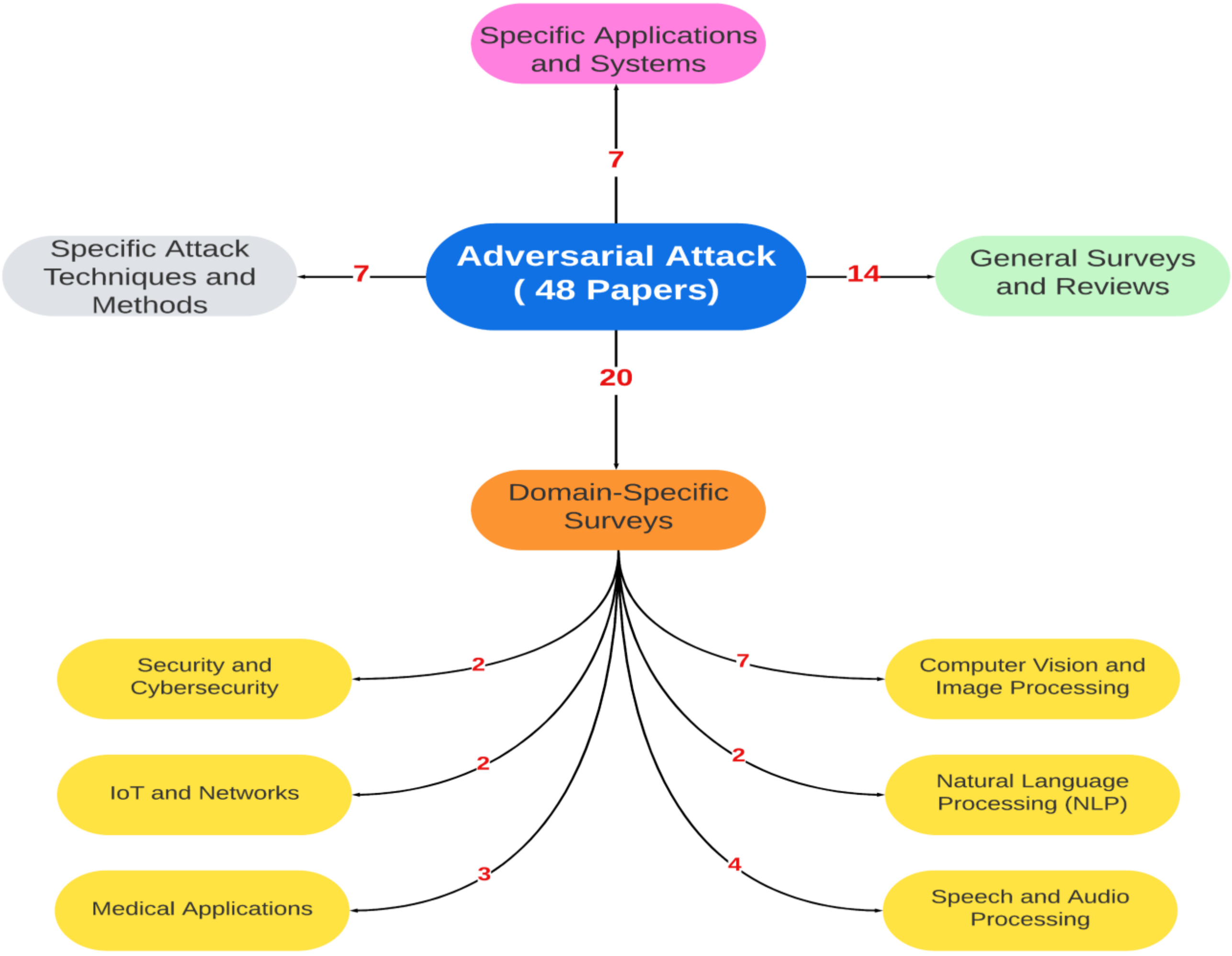
Specific Applications and Systems
7
Specific Attack Techniques and Methods
7
Adversarial Attack (48 Papers)
14
General Surveys and Reviews
20
Domain-Specific Surveys
2
Security and Cybersecurity
2
IoT and Networks
3
Medical Applications
7
Computer Vision and Image Processing
2
Natural Language Processing (NLP)
4
Speech and Audio Processing

Fig. 13. The taxonomy of Surveys and Reviews papers in adversarial attack.

The taxonomy of adversarial attacks is structured into four main topics, which are General Surveys and Reviews, Specific Attack Techniques and Methods, Domain-Specific Surveys and Specific Applications and Systems, as illustrate below.

1. General Surveys and Reviews

This division presents number of research papers that gives an overall view of adversarial attacks and protection cut across all AI arenas. These surveys are the equivalent of foundational resources that help the strategic minds in understanding the nature of the threats and strategies of fighting such threats. These papers surely demonstrate that there are diverse challenges confronting defensive mechanisms against adversarial attacks but there still is a significant need to develop and research further to effectively counter these threats. However, through the encouragement of protection research and development of solid mitigation strategies, scientists could reinforce the resilience of AI systems and save them from going wrong when applied in real life.

2. Domain-Specific Surveys

a. Computer Vision, and Image Processing

The field of research deals with finding the system's vulnerabilities in image classification and object detection by using the model if it is applicable, among them is the visual recognition systems assessment which is seeking to look into adversarial attacks' vulnerabilities with regard to supervising those systems against the risks. Even therefore, adjusting defensive systems to tricks of adversarial attacks in computer vision is still tough problem for which we still meet but undoubtedly, we can and must to deal with this gradually [118]–[124].

b. Natural Language Processing (NLP)

In NLP, surveys on attacks of natural language models are also performed by researchers. Text classifiers are another topics of research. These studies highlight the susceptibility of NLP models to adversarial attacks to adversarial attacks and implications mean for strategy like sentiment analysis and text generation. However, that recognizing adversarial vulnerabilities is a good future work, it is still a pending fact to establish aggressive protection strategies the NLP systems in order to continue to stand against the ever changing attack methods [20], [125].

c. Speech and Audio Processing

This part involves recognizing and processing audio systems that are vulnerable to adversarial attacks on the system level. In this scenario, surveys tap into the area where adversarial attacks pose a threat to speech-based AI models, mainly to speaker recognition and emotion recognition systems. The results provide evidence for greater protection competencies that can impede adversarial attacks to the speech and audio generation technology [126]–[129].

d. Security and Cybersecurity

There are lots of risks of adversarial attacks, particularly for intrusion detection, malware detection, etc. The surveys in this category are meant to estimate which route a negative action is done that is aimed to disrupt the cybersecurity structure, and what tactics are applicable for the protection against those threats. Additionally, computer security is becoming increasingly important due to continually emerging cybercrimes which in turn calls for protection mechanisms as well as adaptability of the security systems to meet the challenges being posed by the adversaries [130], [131].

e. IoT and Networks

As the networks of IoT keep escalating, they silt increasingly into adversarial attacks. The surveys are focused on the security consequences of an advisory intrusion to the IoT platforms and show the need for a secure protection strategy as a mean of defending interconnected devices networks. A solid security positioning of interconnected internet of things networks can be attained only through unwavering collaboration to build up devices with the capability to adapt to novel defenses in the midst of hostile atmosphere of dynamic communication systems [132,133].

f. Medical Applications

Adversarial attacks are a widespread problem in healthcare imaging, as biomedical systems may fall victims to the compromised integrity and dependability of health devices. Distribution of polls related to that in this area discover the possibilities to put on the AI system of the medical type under pressure and will emphasize the impact protection mechanisms have on ensuring patient data security and saving the healthcare system from incorrect conclusions. Securing medical typing of adversarial attacks is key towards the trust in the uses of healthcare AI systems that are dependable [134–136].

3. Specific Attack Techniques and Methods

Intelligent about the details of adversarial attack tactics advancement is the central part of the effective protection plan making. This part based on the identification of the individual methods of attacks like evasion attack and poisoning attack among domains of AI. Researchers can gain a deep understanding on the purpose and nature of hostile instances by studying the intricacies of such techniques. That way, they can develop preventive measures to incorporate protection mechanisms in AI systems and thus deprive the antagonists of the chance to succeed in adversarial attacks [137]–[143].

4. Specific Applications and Systems

This part covers attacks are application-specific, or system-level, which break down, an intelligent vehicle, a recommendation system, or the ready-to-use framework to build the graph-based models. The results of these surveys indicate the number of categories of use, where the attacks can be carried out, and emphasize the trend for the individual approach to the creation of the counteractive measures to this problem. In meeting the particularities of the challenges existing in any hostile environment, academics and engineers will come up with customized defensive gears which will preserve the vital systems and networks [144–150].

Overall, the number of citations also differs greatly from one research paper to another due to the disparity in the utilization, importance and acknowledgment that the papers have received in the scholarly society. Moreover, the number of articles published over various themes is not the same, which depicts the differences in the research interest and productivity of scholars across themes. In order to make these differences more comprehensible, three different diagrams has been designed (see Figures 14, 15, and 16). These diagrams depict the changes in the number of citations and the amount of research published, and compare the values based on our classification scheme. This visual aid makes it easier to do the comparison and also illustrates the rich and varied nature of research output in our area of focus.

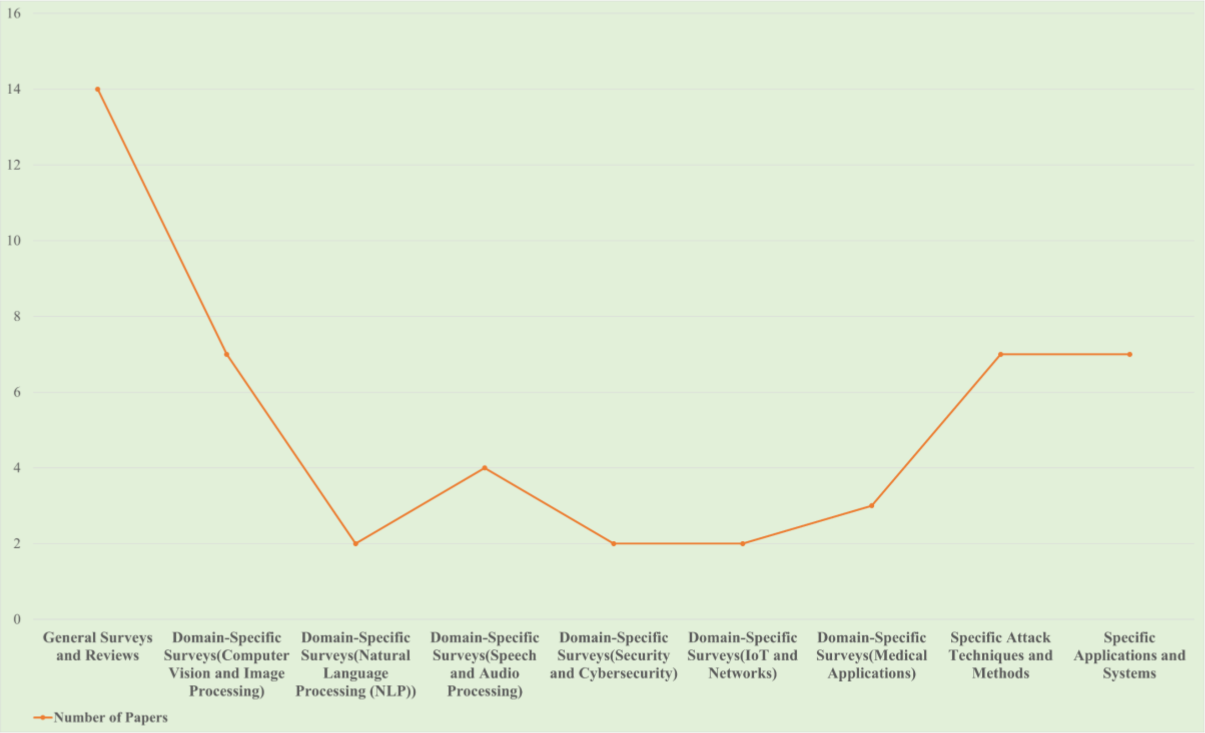

Fig. 14. Number of review papers according to their taxonomy topics.

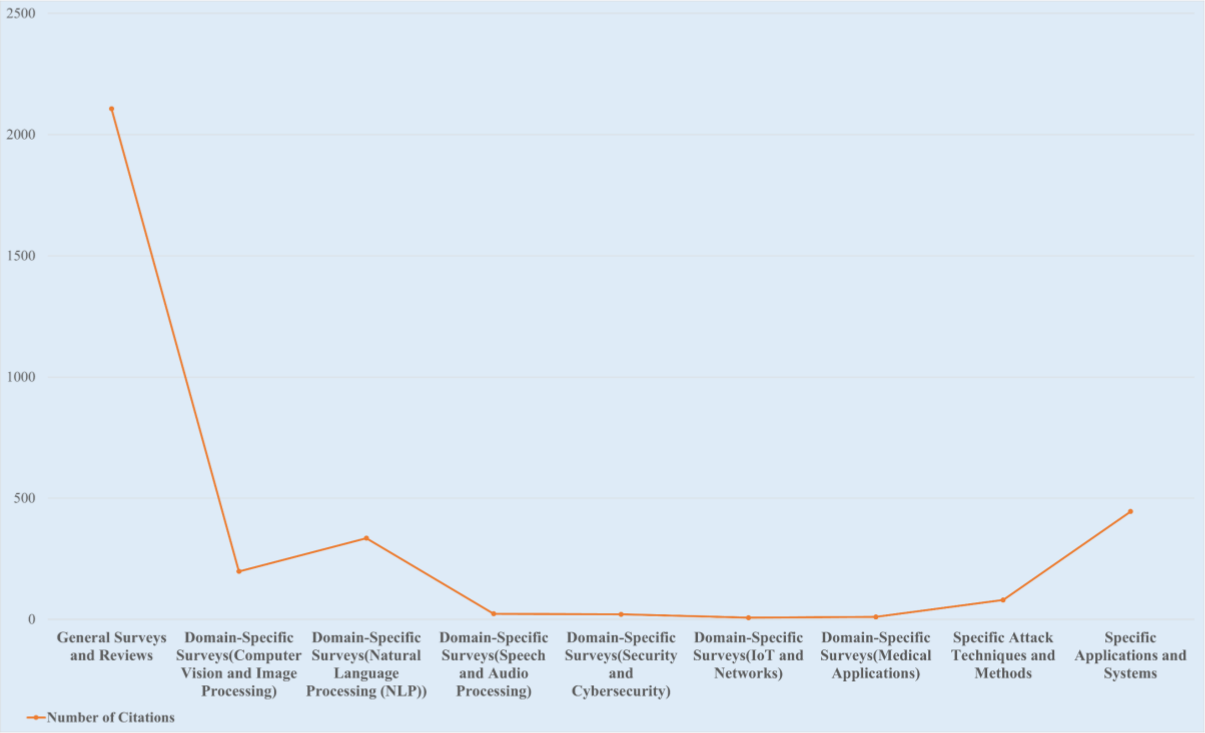

Fig. 15. Number of citations for review papers according to their taxonomy topics.

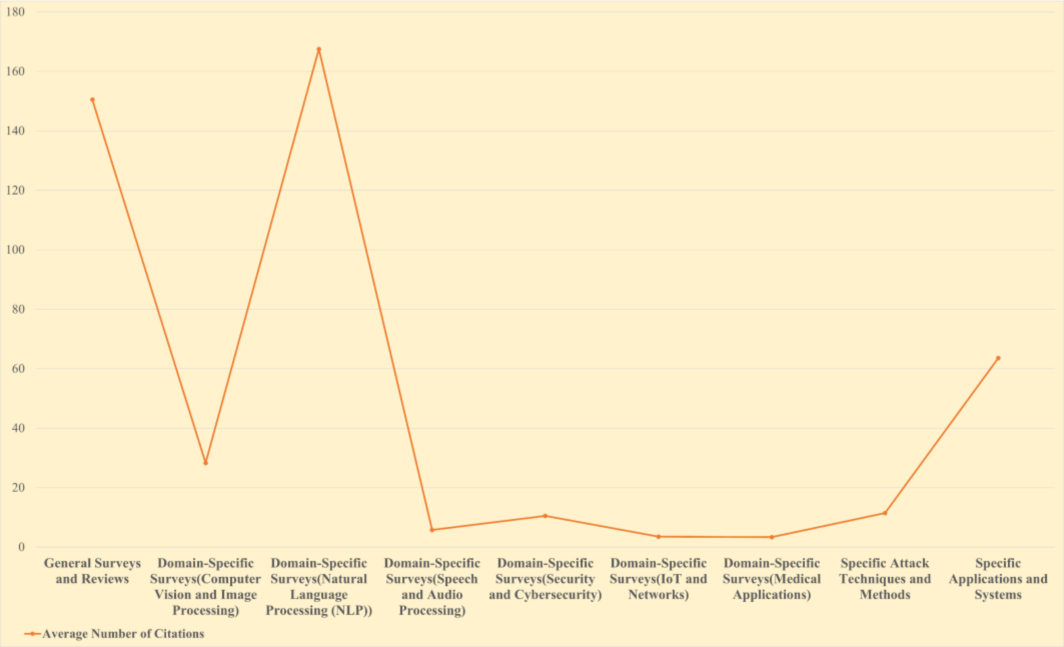

Fig. 16. Average number citations for review papers according to their taxonomy topics.

All in all, scrutinizing further it could be stated that basic strategic approaches to defensive protection in relation to adversarial attacks are ingrained in adversarial or espionage threats. As such, the bulk of the research in this domain can be categorized into three primary areas; the identification of new types of adversarial attacks that may be performed on AI models, the discovery of new techniques used to defend AI models against such attacks, and the improvement of already-known defense techniques. This tripartite focus underscores the central theme around which most research in this field revolves. That is why we can talk about specific aspects as the protection of AI models.

As protection is best known to be of paramount significance in defending the fundamental and formidable nature of artificial intelligence systems against adversarial attacks, this research set to conduct a comprehensive literature review of all noteworthy aspects of protection strategies against adversarial attacks. In this paper, we will analyze potential approaches and solutions that have been proposed to mitigate threat and risks to machine learning and deep learning algorithms. In assessing the current status of these protective measures, we will also review and reflect upon the literature about the theoretical contributions and the actual implementation.

In addition, this research will cover new advances and enhancements in protection techniques, how adversarial attacks are developed against the protection mechanisms, and how generation of protection strategy affects protection mechanisms' robustness. We will also investigate proactive and reactive approaches to Malware-invoked intrusion detection systems as well as their applicability to adversarial machine learning.

The third important element on our review include guidelines on assessing datasets used in adversarial research. This will be followed by a discussion on the availability and quality of these datasets and their potential for supporting development and testing of defense objectives. Mechanizing reasoning necessarily provides a deeper understanding of the types of data available and their potential impact on the research front in order to build efficient and reliable defenses.

Lastly, it is essential to underline that, based on the-centricity of the issue, it possibly can be mentioned that the primary axis which is usually under consideration 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 the effectiveness in terms of protecting the ML/DL models. Thus, the mission of this work is to make a small but significant step towards 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.

4. GAPS ANALYSIS AND OPEN ISSUES

Even though there has been progress in the area of theory and practical models for adversarial attack and defense, there are still some issues which remain open in this area. These are revealed in this section together with identifying potential novel and imaginative measures that could improve the stability as well as the security of AI systems.

4.1 Gaps

a. **Limited Understanding of Attack Mechanisms:** The steps that attackers employ to intimidate artificial intelligence are not examined comprehensively. A few works have explored the specific conditions for such attacks, but there is no comprehensively understanding of how these attacks affect the decision boundaries and what would happen to other types of models.

b. **Inadequate Defense Strategies:** Most of the current state-of-art defense mechanisms are not very powerful. It is also important to mention that the existing protocols may target particular kinds of threats and do not elaborate the sufficient level of protection against all existing types of attacks and approaches in model construction.

c. **Real-world Applicability:** This implies that while theoretical work on this subject is abundant, few scholars address the practical implications or findings in real life. Many of theories that have been developed in relation to these ideas have not been calibrated in actual, raw situations thus grey areas arise as pertains the efficiency of these concepts specifically in the pragmatic setting such as healthcare and fully automated driving.

d. **The many approaches of putting up adversary examples:** It is essential to emphasize that the process of transferability, in which adversarial examples created for one model work with others, is not yet fully explained. This lack of information harms the ability to create global defense structures that would prevent such attacks from being transferred.

e. **Absence of a Standard, and Basic Criteria:** Lack of shared reference points for measuring the effectiveness of evaluating the attacks and, in other words, comparing different solutions to counter the adversarial threats also remains a problem.

4.2 Open Issues

a. **Dynamic Nature of Adversarial Threats:** The threats are not constant and are dynamic therefore it is rather inefficient to be static when it comes to defense. This particular aspect means that there is always a need to update and raise the effectiveness of the existing and ongoing defense mechanisms.

b. **Scalability of Defense Mechanisms:** There are some of the present measures for defense that do not work well especially when used on large databases or large models. With these defenses come their complexity and the overheads that are a major disadvantage in applying the mentioned strategies in real life.

c. **Ethical and Legal Implications:** Adversarial attacks and defenses have certain ethical and legal implications given the fact that the fields such as healthcare and autonomous systems are critical. It also matters to elaborate the threats to the privacy, safety and the legal issues which are also the consequences of those threats.

d. **Cross-domain Vulnerabilities:** One of the possible directions is the analysis of the effects of adversarial attacks in various fields, including Computer vision, NLP, cybersecurity, and the connection between these fields. At the current stage, the study shows that the approach can only work in some fields and the overall impact of the approach has not been investigated.

e. **Human-in-the-loop Systems:** A rather intriguing field is the human in the loop for the identification and countermeasures against adversarial attacks in the context of machine learning. For this reason, it is crucial to design proper and efficient human-in-the-loop systems for the management of the level of automation and human control when formulating adequate defense mechanisms.

4.3 Innovative Key Solutions

a. **Adversarial Training:** The Adversarial training which suggest that models are trained with adversarial examples in addition to the normal data is one of the most promising methodologies. It is utilized to increase the robustness of models against adversarial, in other words, to teach a model not to be deceived by adversarial perturbation.

b. **Ensemble Methods:** Where several models are put together to make a prediction and work on different principles perhaps less likely to be attacked by adversaries. The general working principle of an ensemble which relies on multiple models is rooted in the fact that the models are different and as such it will be unlikely that it would be possible to deceive all of them at the same instance.

c. **Generative Adversarial Networks:** The GANs can be utilized to generate adversarial example that will improve the training of the defensive models.

d. **Adversarial Detection Systems:** Procedures to identify adversarial attacks intended to be in place before their impact is incorporated into the primary model can be useful as the front-line defense. The above-mentioned detection systems can Learn that there are adversarial manipulations in input data and eject such inputs.

.

5. ETHICAL AND LEGAL CONSIDERATIONS

Adversarial attacks in machine learning are not mere technicalities but pose serious concern to aspects of ethical and legal concern such as aspects to do with trust [151]–[153]. Applications of AI and ML systems are integrated into virtually all spheres of life and work including personal and endurable health, financial services, autonomous motor vehicle systems, national security systems and others [154]–[157]; the more these systems are implicated in important activities, the higher the probability of an adversarial hack [158]–[160]. It is with these attacks that the credibility of the verdicts passed by the AI systems can be altered for the worse not only in terms of the technology applied, but also results. For instance, an extensive adversarial attack on a healthcare AI system may result in a wrong assessment of a patient and that is most probable to cause the demise of the involved patient. In the same manner, advanced hacking in self-driving cars could lead to an incident that gives rise to questions of prima facie negligence and corresponding questions of compensation in relation to the occurrence. Consequently, the extension of the use of AI systems that can easily be commandeered calls for significant liability exposures especially to the developers and implementing organizations especially in sensitive areas of operation. Moreover, risks of this sort are magnified by the fact that rules and guidelines are still not well developed around AI, on account of the subject area being still quite new and also because AI is not commonly used in many industries making it unclear how certain existing laws can apply to AI. What is more, the technique used in adversarial techniques themselves can be considered unethical: but they can be specifically designed to deceive people for instance when influencing their opinions or obtaining information while in cyberspace spy craft. Thus, today it is high time that there are good and proper ethical standards and/or laws that will enacted to rightly govern, protect and regulate the deployment and implementation of the models and systems, and especially against adversarial attacks and other. The rise of adversarial machine learning therefore needs the engagement of technologist's ethicists legal professionals and policymakers to come up with a macro approach towards the technical and social processes involved in adversarial machine learning.

6. TRENDS IN FUTURE RESEARCH OF ADVERSARIAL ATTACK

Where AI and ML are rapidly growing, the threats created from adversarial attacks on AI and the defences against them are also constant undergoing changes. The future of research in this area can be predicted on the basis of several trends that appear to be progressing hand in hand with the advances in adversarial techniques and applications of AI systems in vitals areas. Quite possibly one of the most obvious trends would be the increase in the sophistication, or as it were, nuances of the adversarial attacks. Today the majority of the attacks to machine learning models are based on simple manipulations of inputs what may change in the future since the state-of-art allows for the creation of more sophisticated attacks such as using generative models to generate adversarial examples that are indistinguishable from ones that were not manipulated. Such attacks may target the risk in the individual model as well as seek to find systemic vulnerabilities that can affect whole AI ecosystems, which may involve fanning out coordinated multi-vector attacks that can strike at different levels of an application or a system at the same time.

As such, with the change in these threats in the future, it is believed that the formulation of advanced and sophisticated security measures will be a major area of research study. That is, adversarial training is one of the methods under which models are recurrently trained with new adversarial instances. However, a downside is that this kind of approach is difficult to scale – especially as models are elaborated, and the volume of information increases. Subsequent studies will probably look into other strategies of increasing efficiency in adversarial training and apply the technique on a large scale through the help of federated learning or other forms of distributed computation in order not to share sensitive information. Another strong emerging tendencies are connected with usage of explain ability and interpretability in defence strategies. With adversarial attacks growing more complex, it means that any AI solution has not only to minimize the impact of such attacks but also to explain its actions to the user. This will assist to raise confidence in AI systems and reduce the likelihood or risk of an attacker taking advantage of the weakness in AI systems and bringing catastrophic results especially in areas where the use of AI is sensitive such as the medical field, financial sectors and self-driving vehicles.

There is also likely to be much interest in adversarial machine learning and its relation to new technologies. When AI is introduced into new and more various professions, it means computer attacks also occur in new and various professions and in various forms, which also opens up new possibilities for their protection. For instance, probative applications of quantum machine learning may give rise to varieties of adversarial perturbations previously unimagined, which then require new quantum immunizations to deter. In the same regard, the implementation of blockchain technology to protect the AI system may present new ways of creating the guarantee and traceability of data; however, this may also be vulnerable to new types of attack. The Metaverse, the field that is developing at the intersection of virtual reality, augmented reality, and AI, is also especially dangerous for adversarial attacks since getting into the material world is possible there. Research in these areas will have to be highly multidisciplinary in cooperation with experts from the fields such as computer science, cryptography, physics and others.

In the same vein, cross-domain transferability of adversarial attacks can be seen as another direction that is likely to attract more and more focus. Recent studies have shown that, for example, adversarial examples that may be generated to deceive one model can be effective with other models, even though the models are different, or perhaps have been trained using different datasets. Future work will probably focus on broadening the understanding of how this transferability can be achieved with the intention of building generalized defensive strategies that would work on a variety of other kinds of attacks. This could in turn lead to the development of AI models which are, in general, immune to adversarial deformation, irrespective of their environment of use.

Another nerve of the future is a continuously growing interest in the ethical and social aspects of adversarial attacks and their counterparts. Where AI systems are used in increasingly crucial roles in areas like criminology, employment, and police work, there is an increasing realisation that adversarial attacks can either reinforce known sorts of biases or even invent new prejudicial techniques. Future research will have to respond to these challenges by creating defense mechanisms which would be effective from the technical point of view and totally ethical. This may mean coming up with methods on how fair the AI systems can be when attacks are made on them, or constructing new laws and policies for the use of AI in areas with high risk.

Lastly, as for the future of adversarial researches, they will further develop as multidisciplinary and cross-sectoral processes. Since the problems arising from adversarial attacks are becoming more intricate, the role of researchers will be inseparable, and information exchange between academia, businesses, and governments will be vital. These could include setting up new research collaborations and funding partnerships to work on adversarial defence, making new open source tool and data sets available for making progress in this area. The AI community needs to organize its research effort like a large-scale scientific research project addressing the existing and emerging threats and making sure that the new technology, which is AI, is created, used, and implemented safely and ethically

7. CONCLUSION

The area of machine learning adversarial attacks, can be classified as one of the fields which both have great threats along with the great opportunities for its further development. The research of the present era has come to a certain extent in perceiving such weaknesses and the developmental of some primitive forms of safeguarding mechanisms. However, several critical gaps in knowledge are pointed out regarding the attack strategies, efficacy of the protection measures, and the practical application of the suggestions.

What might be done in such attempts is useful for a broader development of safe and stable machine learning models in the future. Learning more regarding the creation of adversarial examples transferability, the creation of unified assessment criteria, and how the creation of successful, dynamically defendable methods of scalability are the next big steps. However, other aspects like ethical and legal perspective of adversarial attacks, inter and intra domain vulnerabilities, and human in the loop in Machine Learning based systems have to explore further. This is so because the new solutions such as adversarial training, the use of ensemble methods, GAN for defence, robustness certification, automated defence, adversarial detectors, and better explainability are among the promising directions. All these approaches do not only improve the robustness of the machine learning models against adversarial attacks but also positively affects the development of the subject to some extent.

Thus, the existing approach to the problem in the current research includes tendencies in the field of studies, must aim at the further and present studies. Thus, different machine learning methods that target adversarial procedures and guarantee the security of AI implementations in various significant areas can be developed to ensure the improvement of stronger, multifunctional, and invulnerable models

Conflicts of Interest

The author's disclosure statement confirms the absence of any conflicts of interest.

Funding

The author's paper clearly indicates that the research was conducted without any funding from external sources.

Acknowledgment

The author extends appreciation to the institution for their unwavering support and encouragement during this research

References

[1] A. S. Albahri *et al.*, "A systematic review of trustworthy artificial intelligence applications in natural disasters," *Comput. Electr. Eng.*, vol. 118, p. 109409, 2024, doi: 10.1016/j.compeleceng.2024.109409.

[2] M. A. Habeeb, "Hate Speech Detection using Deep Learning," Master's thesis, University of Miskolc, Miskolc, Hungary, 2021. [Online]. Available: http://midra.uni-miskolc.hu/document/40792/38399.pdf Accessed: Oct. 28, 2024.

[3] M. E. Alqaysi, A. S. Albahri, and R. A. Hamid, "Evaluation and benchmarking of hybrid machine learning models for autism spectrum disorder diagnosis using a 2-tuple linguistic neutrosophic fuzzy sets-based decision-making model," *Neural Comput. Appl.*, 2024, doi: 10.1007/s00521-024-09905-6.

[4] A. H. Alamoodi, M. S. Al-Samarraay, O. S. Albahri, M. Deveci, A. S. Albahri, and S. Yussof, "Evaluation of energy economic optimization models using multi-criteria decision-making approach," *Expert Syst. Appl.*, vol. 255, p. 124842, 2024, doi: 10.1016/j.eswa.2024.124842.

[5] A. S. Albahri *et al.*, "Prioritizing complex health levels beyond autism triage using fuzzy multi-criteria decision-making," *Complex Intell. Syst.*, 2024, doi: 10.1007/s40747-024-01432-0.

[6] S. Dadvandipour and Y. L. Khaleel, "Application of deep learning algorithms detecting fake and correct textual or verbal news," *Prod. Syst. Inf. Eng.*, vol. 10, no. 2, pp. 37–51, 2022, doi: 10.32968/psaie.2022.2.4.

[7] M. A. Habeeb, Y. L. Khaleel, and A. S. Albahri, "Toward Smart Bicycle Safety: Leveraging Machine Learning Models and Optimal Lighting Solutions," in *Proceedings of the Third International Conference on Innovations in Computing Research (ICR'24)*, K. Daimi and A. Al Sadoon, Eds., Cham: Springer Nature Switzerland, 2024, pp. 120–131.

[8] S. Ghazal, A. Munir, and W. S. Qureshi, "Computer vision in smart agriculture and precision farming: Techniques and applications," *Artif. Intell. Agric.*, vol. 13, pp. 64–83, 2024, doi: 10.1016/j.aiia.2024.06.004.

[9] Y. L. Khaleel, "Fake News Detection Using Deep Learning," Master's thesis, University of Miskolc, Miskolc, Hungary, 2021, doi: 10.1007/978-3-030-91305-2_19.

[10] Z. T. Al-qaysi, A. S. Albahri, M. A. Ahmed, and M. M. Salih, "Dynamic decision-making framework for benchmarking brain–computer interface applications: a fuzzy-weighted zero-inconsistency method for consistent weights and VIKOR for stable rank," *Neural Comput. Appl.*, vol. 36, no. 17, pp. 10355–10378, 2024, doi: 10.1007/s00521-024-09605-1.

[11] A. S. Albahri *et al.*, "Fuzzy decision-making framework for explainable golden multi-machine learning models for real-time adversarial attack detection in Vehicular Ad-hoc Networks," *Inf. Fusion*, vol. 105, p. 102208, 2024, doi: 10.1016/j.inffus.2023.102208.

[12] A. H. Alamoodi *et al.*, "Selection of electric bus models using 2-tuple linguistic T-spherical fuzzy-based decision-making model," *Expert Syst. Appl.*, vol. 249, p. 123498, 2024, doi: https://doi.org/10.1016/j.eswa.2024.123498.

[13] F. K. H. Mihna, M. A. Habeeb, Y. L. Khaleel, Y. H. Ali, and L. A. E. Al-Saeedi, "Using Information Technology for Comprehensive Analysis and Prediction in Forensic Evidence," *Mesopotamian J. CyberSecurity*, vol. 4, no. 1, pp. 4–16, 2024, doi: 10.58496/MJCS/2024/002.

[14] [14] A. S. Albahri, Y. L. Khaleel, and M. A. Habeeb, "The Considerations of Trustworthy AI Components in Generative AI; A Letter to Editor," *App l. Data Sci. Anal.*, vol. 2023, pp. 108–109, 2023, doi: 10.58496/adsa/2023/009.

[15] L. Alzubaidi *et al.*, "MEFF – A model ensemble feature fusion approach for tackling adversarial attacks in medical imaging," *Intell. Syst. with Appl.*, vol. 22, 2024, doi: 10.1016/j.iswa.2024.200355.

[16] D. Zügner and S. Günnemann, "Adversarial attacks on graph neural networks via meta learning," in *7th International Conference on Learning Representations, ICLR 2019*, International Conference on Learning Representations, ICLR, 2019. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-

85083950290&partnerID=40&md5=4a963cbdc0747de26316af0093ca9ad3

[17] O. Ibitoye, O. Shafiq, and A. Matrawy, “Analyzing adversarial attacks against deep learning for intrusion detection in IoT networks,” in *Proceedings - IEEE Global Communications Conference, GLOBECOM*, 2019. doi: 10.1109/GLOBECOM38437.2019.9014337.

[18] M. Haghighatlari, J. Li, F. Heidar-Zadeh, Y. Liu, X. Guan, and T. Head-Gordon, “Learning to Make Chemical Predictions: The Interplay of Feature Representation, Data, and Machine Learning Methods,” *Chem*, vol. 6, no. 7, pp. 1527–1542, 2020, doi: 10.1016/j.chempr.2020.05.014.

[19] J. G. Greener, S. M. Kandathil, L. Moffat, and D. T. Jones, “A guide to machine learning for biologists,” *Nat. Rev. Mol. Cell Biol.*, vol. 23, no. 1, pp. 40–55, 2022, doi: 10.1038/s41580-021-00407-0.

[20] W. E. Zhang, Q. Z. Sheng, A. Alhazmi, and C. Li, “Adversarial Attacks on Deep-learning Models in Natural Language Processing,” *ACM Trans. Intell. Syst. Technol.*, vol. 11, no. 3, pp. 1–41, 2020, doi: 10.1145/3374217.

[21] X. Ma *et al.*, “Understanding adversarial attacks on deep learning based medical image analysis systems,” *Pattern Recognit.*, vol. 110, p. 107332, 2021, doi: 10.1016/j.patcog.2020.107332.

[22] M. Shen, H. Yu, L. Zhu, K. Xu, Q. Li, and J. Hu, “Effective and Robust Physical-World Attacks on Deep Learning Face Recognition Systems,” *IEEE Trans. Inf. Forensics Secur.*, vol. 16, pp. 4063–4077, 2021, doi: 10.1109/TIFS.2021.3102492.

[23] K. Roshan and A. Zafar, “Black-box adversarial transferability: An empirical study in cybersecurity perspective,” *Comput. Secur.*, vol. 141, 2024, doi: 10.1016/j.cose.2024.103853.

[24] K. Roshan, A. Zafar, and S. B. Ul Haque, “Untargeted White-box Adversarial Attack with Heuristic Defence Methods in Real-time Deep Learning based Network Intrusion Detection System,” *Comput. Commun.*, 2023, doi: https://doi.org/10.1016/j.comcom.2023.09.030.

[25] F. Yu, L. Wang, X. Fang, and Y. Zhang, “The defense of adversarial example with conditional generative adversarial networks,” *Secur. Commun. Networks*, vol. 2020, no. 1, p. 3932584, 2020, doi: 10.1155/2020/3932584.

[26] H. Liang, E. He, Y. Zhao, Z. Jia, and H. Li, “Adversarial Attack and Defense: A Survey,” *Electron.*, vol. 11, no. 8, 2022, doi: 10.3390/electronics11081283.

[27] A. Chakraborty, M. Alam, V. Dey, A. Chattopadhyay, and D. Mukhopadhyay, "A survey on adversarial attacks and defences," *CAAI Trans. Intell. Technol.*, vol. 6, no. 1, pp. 25–45, 2021, doi: 10.1049/cit2.12028.

[28] K. Mahmood, R. Mahmood, E. Rathbun, and M. Van Dijk, “Back in Black: A Comparative Evaluation of Recent State-Of-The-Art Black-Box Attacks,” *IEEE Access*, vol. 10, pp. 998–1019, 2022, doi: 10.1109/ACCESS.2021.3138338.

[29] Y. Dong, S. Cheng, T. Pang, H. Su, and J. Zhu, “Query-Efficient Black-Box Adversarial Attacks Guided by a Transfer-Based Prior,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 12, pp. 9536–9548, 2022, doi: 10.1109/TPAMI.2021.3126733.

[30] G. Yang, Q. Ye, and J. Xia, “Unbox the black-box for the medical explainable AI via multi-modal and multi-centre data fusion: A mini-review, two showcases and beyond,” *Inf. Fusion*, vol. 77, pp. 29–52, 2022, doi: 10.1016/j.inffus.2021.07.016.

[31] E. Mariotti, “A holistic perspective on designing and evaluating explainable AI models: from white-box additive models to post-hoc explanations for black-box models.” 2024.

[32] S. Ai, A. S. Voundi Koe, and T. Huang, “Adversarial perturbation in remote sensing image recognition,” *Appl. Soft Comput.*, vol. 105, p. 107252, 2021, doi: 10.1016/j.asoc.2021.107252.

[33] C. Zhang, X. Costa-Perez, and P. Patras, “Adversarial Attacks Against Deep Learning-Based Network Intrusion Detection Systems and Defense Mechanisms,” *IEEE/ACM Trans. Netw.*, vol. 30, no. 3, pp. 1294–1311, 2022, doi: 10.1109/TNET.2021.3137084.

[34] B. Wu *et al.*, “Attacking Adversarial Attacks as A Defense,” *arXiv Prepr. arXiv2106.04938*, 2021, [Online]. Available: http://arxiv.org/abs/2106.04938

[35] N. Liu, M. Du, R. Guo, H. Liu, and X. Hu, “Adversarial Attacks and Defenses,” *ACM SIGKDD Explor. Newsl.*,

vol. 23, no. 1, pp. 86–99, May 2021, doi: 10.1145/3468507.3468519.

[36] L. Griffin, "Evaluating Methods for Improving DNN Robustness Against Adversarial Attacks," no. 1. University of South Florida, pp. 1–23, 2023. [Online]. Available: https://www.proquest.com/openview/0a3e9e510f3b25b0516f4b623af4423f/1?pq-origsite=gscholar&cbl=18750&diss=y

[37] Y. L. Khaleel, M. A. Habeeb, A. S. Albahri, T. Al-Quraishi, O. S. Albahri, and A. H. Alamoodi, "Network and cybersecurity applications of defense in adversarial attacks: A state-of-the-art using machine learning and deep learning methods," J. Inf. Secur, vol. 33, no. 1, 2024, doi: 10.1515/jisys-2024-0153.

[38] M. Macas, C. Wu, and W. Fuertes, "Adversarial examples: A survey of attacks and defenses in deep learning-enabled cybersecurity systems," *Expert Syst. Appl.*, vol. 238, p. 122223, Mar. 2024, doi: 10.1016/j.eswa.2023.122223.

[39] K. Ren, T. Zheng, Z. Qin, and X. Liu, "Adversarial Attacks and Defenses in Deep Learning," *Engineering*, vol. 6, no. 3, pp. 346–360, 2020, doi: 10.1016/j.eng.2019.12.012.

[40] J. Chen, X. Wu, Y. Guo, Y. Liang, and S. Jha, "Towards Evaluating the Robustness of Neural Networks Learned By Transduction," in *ICLR 2022 - 10th International Conference on Learning Representations*, Ieee, 2022, pp. 39–57.

[41] A. Ilyas, S. Santurkar, D. Tsipras, L. Engstrom, B. Tran, and A. Madry, "Adversarial examples are not bugs, they are features," *Adv. Neural Inf. Process. Syst.*, vol. 32, 2019.

[42] X. Zhang and M. Zitnik, "GNNGUARD: Defending graph neural networks against adversarial attacks," in *Advances in Neural Information Processing Systems*, Neural information processing systems foundation, 2020. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85104609818&partnerID=40&md5=7112b4afbafe82c4d30a023c4eeb8dc3

[43] H. Wu, C. Wang, Y. Tyshetskiy, A. Docherty, K. Lu, and L. Zhu, "Adversarial examples for graph data: Deep insights into attack and defense," in *IJCAI International Joint Conference on Artificial Intelligence*, K. S., Ed., International Joint Conferences on Artificial Intelligence, 2019, pp. 4816–4823. doi: 10.24963/ijcai.2019/669.

[44] H. Dai *et al.*, "Adversarial attack on graph structured data," in *35th International Conference on Machine Learning, ICML 2018*, K. A. and D. J., Eds., International Machine Learning Society (IMLS), 2018, pp. 1799–1808. [Online]. Available:https://www.scopus.com/inward/record.uri?eid=2-s2.0 85057260187&partnerID=40&md5=f27682ae15830c7b87e1ffcf27e737aa

[45] D. Zügner, A. Akbarnejad, and S. Günnemann, "Adversarial attacks on neural networks for graph data," in *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Association for Computing Machinery, 2018, pp. 2847–2856. doi: 10.1145/3219819.3220078.

[46] Y. Dong *et al.*, "Efficient decision-based black-box adversarial attacks on face recognition," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, IEEE Computer Society, 2019, pp. 7706–7714. doi: 10.1109/CVPR.2019.00790.

[47] S. G. Finlayson, J. D. Bowers, J. Ito, J. L. Zittrain, A. L. Beam, and I. S. Kohane, "Adversarial attacks on medical machine learning," *Science (80-.).*, vol. 363, no. 6433, pp. 1287–1289, 2019, doi: 10.1126/science.aaw4399.

[48] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," in *6th International Conference on Learning Representations, ICLR 2018 - Conference Track Proceedings*, International Conference on Learning Representations, ICLR, 2018. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85083954061&partnerID=40&md5=84bf66031966d7b8f24a2260e59ff64c

[49] R. Shao, X. Lan, J. Li, and P. C. Yuen, "Multi-adversarial discriminative deep domain generalization for face presentation attack detection," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, IEEE Computer Society, 2019, pp. 10015–10023. doi: 10.1109/CVPR.2019.01026.

[50] M. Sadeghi and E. G. Larsson, "Adversarial attacks on deep-learning based radio signal classification," *IEEE Wirel. Commun. Lett.*, vol. 8, no. 1, pp. 213–216, 2019, doi: 10.1109/LWC.2018.2867459.

[51] L. Schönherr, K. Kohls, S. Zeiler, T. Holz, and D. Kolossa, "Adversarial Attacks Against Automatic Speech

Recognition Systems via Psychoacoustic Hiding," in *26th Annual Network and Distributed System Security Symposium, NDSS 2019*, The Internet Society, 2019. doi: 10.14722/ndss.2019.23288.

[52] G. Goswami, N. Ratha, A. Agarwal, R. Singh, and M. Vatsa, "Unravelling robustness of deep learning based face recognition against adversarial attacks," in *32nd AAAI Conference on Artificial Intelligence, AAAI 2018,* AAAI press, 2018, pp. 6829–6836. doi: 10.1609/aaai.v32i1.12341.

[53] S. Huang, N. Papernot, I. Goodfellow, Y. Duan, and P. Abbeel, "Adversarial attacks on neural network policies," in *5th International Conference on Learning Representations, ICLR 2017 - Workshop Track Proceedings*, International Conference on Learning Representations, ICLR, 2017. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85115871809&partnerID=40&md5=84171d54c437457e576e50692d1b7e04

[54] M. Yu, M. Zhou, and W. Su, "A secure routing protocol against byzantine attacks for MANETs in adversarial environments," *IEEE Trans. Veh. Technol.*, vol. 58, no. 1, pp. 449–460, 2009, doi: 10.1109/TVT.2008.923683.

[55] I. Corona, G. Giacinto, and F. Roli, "Adversarial attacks against intrusion detection systems: Taxonomy, solutions and open issues," *Inf. Sci. (Ny).*, vol. 239, pp. 201–225, 2013, doi: 10.1016/j.ins.2013.03.022.

[56] H. Qiu, T. Dong, T. Zhang, J. Lu, G. Memmi, and M. Qiu, "Adversarial Attacks against Network Intrusion Detection in IoT Systems," *IEEE Internet Things J.*, vol. 8, no. 13, pp. 10327–10335, 2021, doi: 10.1109/JIOT.2020.3048038.

[57] S. Cheng, Y. Dong, T. Pang, H. Su, and J. Zhu, "Improving black-box adversarial attacks with a transfer-based prior," in *Advances in Neural Information Processing Systems*, Neural information processing systems foundation, 2019. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85090169867&partnerID=40&md5=9dde3ef3249f74f65c428131ed49734c

[58] Q. Huang, I. Katsman, Z. Gu, H. He, S. Belongie, and S. N. Lim, "Enhancing adversarial example transferability with an intermediate level attack," in *Proceedings of the IEEE International Conference on Computer Vision*, Institute of Electrical and Electronics Engineers Inc., 2019, pp. 4732–4741. doi: 10.1109/ICCV.2019.00483.

[59] Y. Liu, X. Chen, C. Liu, and D. Song, "Delving into transferable adversarial examples and black-box attacks," in *5th International Conference on Learning Representations, ICLR 2017 - Conference Track Proceedings*, International Conference on Learning Representations, ICLR, 2017. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85088225756&partnerID=40&md5=84bfcdfc146d3ce2983bbd860c3547ce

[60] A. Demontis *et al.*, "Why do adversarial attacks transfer? Explaining transferability of evasion and poisoning attacks," in *Proceedings of the 28th USENIX Security Symposium*, USENIX Association, 2019, pp. 321–338. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85072900690&partnerID=40&md5=ebe100a708426ec0b1edeef53c787f14

[61] S. Thys, W. Van Ranst, and T. Goedeme, "Fooling automated surveillance cameras: Adversarial patches to attack person detection," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, IEEE Computer Society, 2019, pp. 49–55. doi: 10.1109/CVPRW.2019.00012.

[62] N. Narodytska and S. Kasiviswanathan, "Simple Black-Box Adversarial Attacks on Deep Neural Networks," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, IEEE Computer Society, 2017, pp. 1310–1318. doi: 10.1109/CVPRW.2017.172.

[63] X. Zhou, W. Liang, W. Li, K. Yan, S. Shimizu, and K. I. K. Wang, "Hierarchical Adversarial Attacks Against Graph-Neural-Network-Based IoT Network Intrusion Detection System," *IEEE Internet Things J.*, vol. 9, no. 12, pp. 9310–9319, 2022, doi: 10.1109/JIOT.2021.3130434.

[64] F. Croce and M. Hein, "Minimally distorted adversarial examples with a fast adaptive boundary attack," in *37th International Conference on Machine Learning, ICML 2020*, D. H. and S. A., Eds., International Machine Learning Society (IMLS), 2020, pp. 2174–2183. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85104186671&partnerID=40&md5=7a4457680c875323e63ed2b49b1b8402

[65] M. Andriushchenko, F. Croce, N. Flammarion, and M. Hein, "Square Attack: A Query-Efficient Black-Box Adversarial Attack via Random Search," *Lect. Notes Comput. Sci.* (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), vol. 12368 LNCS, pp. 484–501, 2020, doi: 10.1007/978-3-030-58592-1_29.

[66] A. Ilyas, L. Engstrom, and A. Madry, “Prior convictions: Black-box adversarial attacks with bandits and priors,” in *7th International Conference on Learning Representations, ICLR 2019*, International Conference on Learning Representations, ICLR, 2019. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85083953118&partnerID=40&md5=d6ac28d6fc01771e2ac0f95812dae7b7

[67] C. Guo, J. R. Gardner, Y. You, A. G. Wilson, and K. Q. Weinberger, “Simple black-box adversarial attacks,” in *36th International Conference on Machine Learning, ICML 2019*, International Machine Learning Society (IMLS), 2019, pp. 4410 – 4423. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85073203396&partnerID=40&md5=dc3fedb1bf2c908a3b08ba70db8571eb

[68] F. Tramèr, A. Kurakin, N. Papernot, I. Goodfellow, D. Boneh, and P. McDaniel, “Ensemble adversarial training: Attacks and defenses,” in *6th International Conference on Learning Representations, ICLR 2018 - Conference Track Proceedings*, International Conference on Learning Representations, ICLR, 2018. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85083953449&partnerID=40&md5=bf2f71b308c2b1ad53207565cf750d26

[69] F. Croce and M. Hein, “Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks,” in *37th International Conference on Machine Learning, ICML 2020*, D. H. and S. A., Eds., International Machine Learning Society (IMLS), 2020, pp. 2184–2194. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85105183209&partnerID=40&md5=f4d452a7ea209d57876bae3aed272167

[70] P. Samangouei, M. Kabkab, and R. Chellappa, “Defense-Gan: Protecting classifiers against adversarial attacks using generative models,” in *6th International Conference on Learning Representations, ICLR 2018 - Conference Track Proceedings*, International Conference on Learning Representations, ICLR, 2018. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85083952288&partnerID=40&md5=89d3e0934167ea8524dd3ea46d84758e

[71] H. Zhang and J. Wang, “Defense against adversarial attacks using feature scattering-based adversarial training,” in *Advances in Neural Information Processing Systems*, Neural information processing systems foundation, 2019. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85085197234&partnerID=40&md5=ad449f48773632064dc822eef014bb04

[72] S. Chen *et al.*, “Automated poisoning attacks and defenses in malware detection systems: An adversarial machine learning approach,” *Comput. Secur.*, vol. 73, pp. 326–344, 2018, doi: 10.1016/j.cose.2017.11.007.

[73] F. Pierazzi, F. Pendlebury, J. Cortellazzi, and L. Cavallaro, “Intriguing properties of adversarial ML attacks in the problem space,” in *Proceedings - IEEE Symposium on Security and Privacy*, 2020, pp. 1332–1349. doi: 10.1109/SP40000.2020.00073.

[74] N. Entezari, S. A. Al-Sayouri, A. Darvishzadeh, and E. E. Papalexakis, “All you need is Low (rank): Defending against adversarial attacks on graphs,” in *WSDM 2020 - Proceedings of the 13th International Conference on Web Search and Data Mining*, Association for Computing Machinery, Inc, 2020, pp. 169–177. doi: 10.1145/3336191.3371789.

[75] K. Lee, K. Lee, H. Lee, and J. Shin, “A simple unified framework for detecting out-of-distribution samples and adversarial attacks,” in *Advances in Neural Information Processing Systems*, G. K., B. S., W. H., G. K., L. H., C.-B. N., and G. R., Eds., Neural information processing systems foundation, 2018, pp. 7167–7177. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85064822451&partnerID=40&md5=ed7d4489422d4e8f072fb9146ddf8761

[76] A. Pattanaik, Z. Tang, S. Liu, G. Bommannan, and G. Chowdhary, “Robust Deep Reinforcement Learning with adversarial attacks,” in *Proceedings of the International Joint Conference on Autonomous Agents and Multiagent Systems, AAMAS*, International Foundation for Autonomous Agents and Multiagent Systems (IFAAMAS), 2018, pp. 2040–2042. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85054762037&partnerID=40&md5=30fcb05b34d27197cdd5dafbddd136ad

[77] [77] Y. Zhao, H. Zhu, R. Liang, Q. Shen, S. Zhang, and K. Chen, "Seeing isn't believing: Towards more robust adversarial attack against real world object detectors," in *Proceedings of the ACM Conference on Computer and Communications Security*, Association for Computing Machinery, 2019, pp. 1989–2004. doi: 10.1145/3319535.33542.

[78] Z. He, A. S. Rakin, and D. Fan, “Parametric noise injection: Trainable randomness to improve deep neural network robustness against adversarial attack,” in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, IEEE Computer Society, 2019, pp. 588–597. doi: 10.1109/CVPR.2019.00068.

[79] F. Zhang, P. P. K. Chan, B. Biggio, D. S. Yeung, and F. Roli, “Adversarial Feature Selection Against Evasion Attacks,” *IEEE Trans. Cybern.*, vol. 46, no. 3, pp. 766–777, 2016, doi: 10.1109/TCYB.2015.2415032.

[80] P. Y. Chen, Y. Sharma, H. Zhang, J. Yi, and C. J. Hsieh, “EAD: Elastic-net attacks to deep neural networks via adversarial examples,” in *32nd AAAI Conference on Artificial Intelligence, AAAI 2018,* AAAI press, 2018, pp. 10–17. doi: 10.1609/aaai.v32i1.11302.

[81] Y. Dong, T. Pang, H. Su, and J. Zhu, “Evading defenses to transferable adversarial examples by translation-invariant attacks,” in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition,* Jun. 2019, pp. 4307–4316. doi: 10.1109/CVPR.2019.00444.

[82] D. Slack, S. Hilgard, E. Jia, S. Singh, and H. Lakkaraju, “Fooling LIME and SHAP: Adversarial attacks on post hoc explanation methods,” in *AIES 2020 - Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, Association for Computing Machinery, Inc, 2020, pp. 180–186. doi: 10.1145/3375627.3375830.

[83] Y. C. Lin, Z. W. Hong, Y. H. Liao, M. L. Shih, M. Y. Liu, and M. Sun, “Tactics of adversarial attack on deep reinforcement learning agents,” in *5th International Conference on Learning Representations, ICLR 2017 - Workshop Track Proceedings*, International Conference on Learning Representations, ICLR, 2017. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85106092143&partnerID=40&md5=fdb1f5623b6062f9121010a12a700836

[84] Y. C. Lin, Z. W. Hong, Y. H. Liao, M. L. Shih, M. Y. Liu, and M. Sun, “Tactics of adversarial attack on deep reinforcement learning agents,” in *IJCAI International Joint Conference on Artificial Intelligence*, S. C., Ed., International Joint Conferences on Artificial Intelligence, 2017, pp. 3756–3762. doi: 10.24963/ijcai.2017/525.

[85] N. Carlini and D. Wagner, “Audio adversarial examples: Targeted attacks on speech-to-text,” in *Proceedings - 2018 IEEE Symposium on Security and Privacy Workshops, SPW 2018*, Institute of Electrical and Electronics Engineers Inc., 2018, pp. 1–7. doi: 10.1109/SPW.2018.00009.

[86] S. Pawar, S. El Rouayheb, and K. Ramchandran, “Securing dynamic distributed storage systems against eavesdropping and adversarial attacks,” *IEEE Trans. Inf. Theory*, vol. 57, no. 10, pp. 6734–6753, 2011, doi: 10.1109/TIT.2011.2162191.

[87] A. Eyas, L. Engstrom, A. Athalye, and J. Lin, “Black-box adversarial attacks with limited queries and information,” in *35th International Conference on Machine Learning, ICML 2018*, D. J. and K. A., Eds., International Machine Learning Society (IMLS), 2018, pp. 3392–3401. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85055122654&partnerID=40&md5=d545b57a0d5af74d7ecf91c57218924e

[88] Y. Cao *et al.*, “Adversarial sensor attack on LiDAR-based perception in autonomous driving,” in *Proceedings of the ACM Conference on Computer and Communications Security*, Association for Computing Machinery, 2019, pp. 2267–2281. doi: 10.1145/3319535.3339815.

[89] X. Wang and K. He, “Enhancing the Transferability of Adversarial Attacks through Variance Tuning,” in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, IEEE Computer Society, 2021, pp. 1924–1933. doi: 10.1109/CVPR46437.2021.00196.

[90] A. Bojchevski and S. Günnemann, “Adversarial attacks on node embeddings via graph poisoning,” in *36th International Conference on Machine Learning, ICML 2019*, International Machine Learning Society (IMLS), 2019, pp. 1112–1123. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85071150199&partnerID=40&md5=fbaebf786559a6129c43664c9c5cba7e

[91] R. Duan, X. Ma, Y. Wang, J. Bailey, A. K. Qin, and Y. Yang, “Adversarial camouflage: Hiding physical-world attacks with natural styles,” in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition,* IEEE Computer Society, 2020, pp. 997–1005. doi: 10.1109/CVPR42600.2020.00108.

[92] Y. Dong *et al.*, “Boosting Adversarial Attacks with Momentum,” in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2018, pp. 9185–9193. doi: 10.1109/CVPR.2018.00957.

[93] J. Zhang *et al.*, "Attacks which do not kill training make adversarial learning stronger," in *37th International Conference on Machine Learning, ICML 2020*, D. H. and S. A., Eds., International Machine Learning Society (IMLS), 2020, pp. 11214–11224. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85105322490&partnerID=40&md5=6e99952b1c309b7719291c84f91bd689

[94] F. Tramèr, N. Carlini, W. Brendel, and A. Madry, "On adaptive attacks to adversarial example defenses," in *Advances in Neural Information Processing Systems*, Neural information processing systems foundation, 2020. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85100497968&partnerID=40&md5=1a2f0668e1a16348fc1f94bd242ee078

[95] M. S. Chong, M. Wakaiki, and J. P. Hespanha, "Observability of linear systems under adversarial attacks," in *Proceedings of the American Control Conference*, Institute of Electrical and Electronics Engineers Inc., 2015, pp. 2439–2444. doi: 10.1109/ACC.2015.7171098.

[96] Y. Lin, H. Zhao, X. Ma, Y. Tu, and M. Wang, "Adversarial Attacks in Modulation Recognition with Convolutional Neural Networks," *IEEE Trans. Reliab.*, vol. 70, no. 1, pp. 389–401, 2021, doi: 10.1109/TR.2020.3032744.

[97] D. Zhu, P. Cui, Z. Zhang, and W. Zhu, "Robust graph convolutional networks against adversarial attacks," in *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Association for Computing Machinery, 2019, pp. 1399–1407. doi: 10.1145/3292500.3330851.

[98] D. J. Miller, Z. Xiang, and G. Kesidis, "Adversarial Learning Targeting Deep Neural Network Classification: A Comprehensive Review of Defenses against Attacks," *Proc. IEEE*, vol. 108, no. 3, pp. 402–433, 2020, doi: 10.1109/JPROC.2020.2970615.

[99] J. Jia, A. Salem, M. Backes, Y. Zhang, and N. Z. Gong, "Memguard: Defending against black-box membership inference attacks via adversarial examples," in *Proceedings of the ACM Conference on Computer and Communications Security*, Association for Computing Machinery, 2019, pp. 259–274. doi: 10.1145/3319535.3363201.

[100] A. Arnab, O. Miksik, and P. H. S. Torr, "On the robustness of semantic segmentation models to adversarial attacks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 12, pp. 3040–3053, 2020, doi: 10.1109/TPAMI.2019.2919707.

[101] H. Xiao, H. Xiao, and C. Eckert, "Adversarial label flips attack on support vector machines," *Front. Artif. Intell. Appl.*, vol. 242, pp. 870–875, 2012, doi: 10.3233/978-1-61499-098-7-870.

[102] S. Baluja and I. Fischer, "Learning to attack: Adversarial transformation networks," in *32nd AAAI Conference on Artificial Intelligence, AAAI 2018*, AAAI press, 2018, pp. 2687–2695. doi: 10.1609/aaai.v32i1.11672.

[103] J. Zhang, J. Chen, D. Wu, B. Chen, and S. Yu, "Poisoning attack in federated learning using generative adversarial nets," in *Proceedings - 2019 18th IEEE International Conference on Trust, Security and Privacy in Computing and Communications/13th IEEE International Conference on Big Data Science and Engineering, TrustCom/BigDataSE 2019*, Institute of Electrical and Electronics Engineers Inc., 2019, pp. 374–380. doi: 10.1109/TrustCom/BigDataSE.2019.00057.

[104] J. Uesato, B. O'Donoghue, A. Van Den Oord, and P. Kohli, "Adversarial risk and the dangers of evaluating against weak attacks," in *35th International Conference on Machine Learning, ICML 2018*, K. A. and D. J., Eds., International Machine Learning Society (IMLS), 2018, pp. 7995–8007. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85057301673&partnerID=40&md5=53ebff6be03140645e37b4d3a505d486

[105] J. Rony, L. G. Hafemann, L. S. Oliveira, I. Ben Ayed, R. Sabourin, and E. Granger, "Decoupling direction and norm for efficient gradient-based l2 adversarial attacks and defenses," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, IEEE Computer Society, 2019, pp. 4317–4325. doi: 10.1109/CVPR.2019.00445.

[106] H. Fawzi, P. Tabuada, and S. Diggavi, "Secure estimation and control for cyber-physical systems under adversarial attacks," *IEEE Trans. Automat. Contr.*, vol. 59, no. 6, pp. 1454–1467, 2014, doi: 10.1109/TAC.2014.2303233.

[107] S. T. Chen, C. Cornelius, J. Martin, and D. H. P. Chau, "ShapeShifter: Robust physical adversarial attack on faster R-CNN object detector," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 11051 LNAI, pp. 52–68, 2019, doi: 10.1007/978-3-030-10925-7_4.

[108] X. Yuan, P. He, Q. Zhu, and X. Li, "Adversarial Examples: Attacks and Defenses for Deep Learning," *IEEE Trans. neural networks Learn. Syst.*, vol. 30, no. 9, pp. 2805–2824, 2019, doi: 10.1109/TNNLS.2018.2886017.

[109] M. Usama, M. Asim, S. Latif, J. Qadir, and Ala-Al-Fuqaha, "Generative adversarial networks for launching and thwarting adversarial attacks on network intrusion detection systems," in *2019 15th International Wireless Communications and Mobile Computing Conference, IWCMC 2019*, Institute of Electrical and Electronics Engineers Inc., 2019, pp. 78–83. doi: 10.1109/IWCMC.2019.8766353.

[110] Z. Wang, H. Guo, Z. Zhang, W. Liu, Z. Qin, and K. Ren, "Feature Importance-aware Transferable Adversarial Attacks," in *Proceedings of the IEEE International Conference on Computer Vision*, Institute of Electrical and Electronics Engineers Inc., 2021, pp. 7619–7628. doi: 10.1109/ICCV48922.2021.00754.

[111] J. Sun, Y. Cao, Q. A. Chen, and Z. Morley Mao, "Towards robust LiDAR-based perception in autonomous driving: General black-box adversarial sensor attack and countermeasures," in *Proceedings of the 29th USENIX Security Symposium*, USENIX Association, 2020, pp. 877–894. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85091956673&partnerID=40&md5=d2c5942e1a7d98b8faa74b1761e19f25

[112] A. Prakash, N. Moran, S. Garber, A. Dilillo, and J. Storer, "Deflecting Adversarial Attacks with Pixel Deflection," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, IEEE Computer Society, 2018, pp. 8571–8580. doi: 10.1109/CVPR.2018.00894.

[113] W. Brendel, J. Rauber, and M. Bethge, "Decision-based adversarial attacks: Reliable attacks against black-box machine learning models," in *6th International Conference on Learning Representations, ICLR 2018 - Conference Track Proceedings*, International Conference on Learning Representations, ICLR, 2018. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85083954048&partnerID=40&md5=d2646cdf5c872ab2eaf5119b5338131f

[114] J. X. Morris, E. Lifland, J. Y. Yoo, J. Grigsby, D. Jin, and Y. Qi, "TextAttack: A Framework for Adversarial Attacks, Data Augmentation, and Adversarial Training in NLP," in *EMNLP 2020 - Conference on Empirical Methods in Natural Language Processing, Proceedings of Systems Demonstrations*, L. Q. and S. D., Eds., Association for Computational Linguistics (ACL), 2020, pp. 119–126. doi: 10.18653/v1/2020.emnlp-demos.16.

[115] L. Li, R. Ma, Q. Guo, X. Xue, and X. Qiu, "BERT-ATTACK: Adversarial attack against BERT using BERT," in *EMNLP 2020 - 2020 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference*, Association for Computational Linguistics (ACL), 2020, pp. 6193–6202. doi: 10.18653/v1/2020.emnlp-main.500.

[116] F. Liao, M. Liang, Y. Dong, T. Pang, X. Hu, and J. Zhu, "Defense Against Adversarial Attacks Using High-Level Representation Guided Denoiser," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Jun. 2018, pp. 1778–1787. doi: 10.1109/CVPR.2018.00191.

[117] J. Lin, C. Song, K. He, L. Wang, and J. E. Hopcroft, "Nesterov Accelerated Gradient and Scale Invariance for Adversarial Attacks," in *8th International Conference on Learning Representations, ICLR 2020*, International Conference on Learning Representations, ICLR, 2020. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85150590715&partnerID=40&md5=732ac1f089eacc316281df94cc07ea5a

[118] Y. Wang, Y. an Tan, W. Zhang, Y. Zhao, and X. Kuang, "An adversarial attack on DNN-based black-box object detectors," *J. Netw. Comput. Appl.*, vol. 161, 2020, doi: 10.1016/j.jnca.2020.102634.

[119] S. M. K. A. Kazmi, N. Aafaq, M. A. Khan, M. Khalil, and A. Saleem, "From Pixel to Peril: Investigating Adversarial Attacks on Aerial Imagery Through Comprehensive Review and Prospective Trajectories," *IEEE Access*, vol. 11, pp. 81256–81278, 2023, doi: 10.1109/ACCESS.2023.3299878.

[120] J. Fang, Y. Jiang, C. Jiang, Z. L. Jiang, C. Liu, and S. M. Yiu, "State-of-the-art optical-based physical adversarial attacks for deep learning computer vision systems," *Expert Syst. Appl.*, vol. 250, 2024, doi: 10.1016/j.eswa.2024.123761.

[121] T. Long, Q. Gao, L. Xu, and Z. Zhou, "A survey on adversarial attacks in computer vision: Taxonomy, visualization and future directions," *Comput. Secur.*, vol. 121, 2022, doi: 10.1016/j.cose.2022.102847.

[122] X. Ling *et al.*, "Adversarial attacks against Windows PE malware detection: A survey of the state-of-the-art,"

Comput. Secur., vol. 128, 2023, doi: 10.1016/j.cose.2023.103134.

[123] N. Akhtar, A. Mian, N. Kardan, and M. Shah, “Advances in Adversarial Attacks and Defenses in Computer Vision: A Survey,” *IEEE Access*, vol. 9, pp. 155161–155196, 2021, doi: 10.1109/ACCESS.2021.3127960.

[124] C. Li, H. Wang, W. Yao, and T. Jiang, “Adversarial attacks in computer vision: a survey,” *J. Membr. Comput.*, vol. 6, no. 2, pp. 130–147, 2024, doi: 10.1007/s41965-024-00142-3.

[125] H. Zheng *et al.*, “Survey of Adversarial Attack, Defense and Robustness Analysis for Natural Language Processing,” *Jisuanji Yanjiu yu Fazhan/Computer Res. Dev.*, vol. 58, no. 8, pp. 1727–1750, 2021, doi: 10.7544/issn1000-1239.2021.20210304.

[126] H. Tan, L. Wang, H. Zhang, J. Zhang, M. Shafiq, and Z. Gu, “Adversarial Attack and Defense Strategies of Speaker Recognition Systems: A Survey,” *Electron.*, vol. 11, no. 14, 2022, doi: 10.3390/electronics11142183.

[127] C. Wei, M. Sun, X. Zou, and X. Zhang, “Reviews on the Attack and Defense Methods of Voice Adversarial Examples,” *J. Cyber Secur.*, vol. 7, no. 1, pp. 100–113, 2022, doi: 10.19363/J.cnki.cn10-1380/tn.2022.01.07.

[128] J. Gao, D. Yan, and M. Dong, “Black-box adversarial attacks through speech distortion for speech emotion recognition,” *Eurasip J. Audio, Speech, Music Process.*, vol. 2022, no. 1, 2022, doi: 10.1186/s13636-022-00254-7.

[129] D. Xu, R. Fang, B. Jiang, and Q. Xuan, “A Review of Speech Adversarial Attack and Defense Methods,” *J. Cyber Secur.*, vol. 7, no. 1, pp. 126–144, 2022, doi: 10.19363/J.cnki.cn10-1380/tn.2022.01.09.

[130] A. Alotaibi and M. A. Rassam, “Adversarial Machine Learning Attacks against Intrusion Detection Systems: A Survey on Strategies and Defense,” *Futur. Internet*, vol. 15, no. 2, 2023, doi: 10.3390/fi15020062.

[131] J. Li, Y. Wang, L. Luo, and Y. Wang, “A Survey of Adversarial Attack Techniques for Android Malware Detection,” *J. Cyber Secur.*, vol. 6, no. 4, pp. 28–43, 2021, doi: 10.19363/J.cnki.cn10-1380/tn.2021.07.02.

[132] H. Khazane, M. Ridouani, F. Salahdine, and N. Kaabouch, “A Holistic Review of Machine Learning Adversarial Attacks in IoT Networks,” *Futur. Internet*, vol. 16, no. 1, 2024, doi: 10.3390/fi16010032.

[133] J. Vitorino, I. Praça, and E. Maia, “SoK: Realistic adversarial attacks and defenses for intelligent network intrusion detection,” *Comput. Secur.*, vol. 134, 2023, doi: 10.1016/j.cose.2023.103433.

[134] A. M. Zbrzezny and A. E. Grzybowski, “Deceptive Tricks in Artificial Intelligence: Adversarial Attacks in Ophthalmology,” *J. Clin. Med.*, vol. 12, no. 9, 2023, doi: 10.3390/jcm12093266.

[135] G. W. Muoka *et al.*, “A Comprehensive Review and Analysis of Deep Learning-Based Medical Image Adversarial Attack and Defense,” *Mathematics*, vol. 11, no. 20, 2023, doi: 10.3390/math11204272.

[136] V. Sorin, S. Soffer, B. S. Glicksberg, Y. Barash, E. Konen, and E. Klang, “Adversarial attacks in radiology – A systematic review,” *Eur. J. Radiol.*, vol. 167, 2023, doi: 10.1016/j.ejrad.2023.111085.

[137] V. Srinivasan, C. Rohrer, A. Marban, K. R. Müller, W. Samek, and S. Nakajima, “Robustifying models against adversarial attacks by Langevin dynamics,” *Neural Networks*, vol. 137, pp. 1–17, 2021, doi: 10.1016/j.neunet.2020.12.024.

[138] M. Xia, Z. Ye, W. Zhao, R. Yi, and Y. Liu, “Adversarial attack and interpretability of the deep neural net- work from the geometric perspective,” *Sci. Sin. Informationis*, vol. 51, no. 9, pp. 1411–1437, 2021, doi: 10.1360/SSI-2020-0169.

[139] X. Han, Y. Zhang, W. Wang, and B. Wang, “Text Adversarial Attacks and Defenses: Issues, Taxonomy, and Perspectives,” *Secur. Commun. Networks*, vol. 2022, 2022, doi: 10.1155/2022/6458488.

[140] A. Kloukiniotis, A. Papandreou, A. Lalos, P. Kapsalas, D. V. Nguyen, and K. Moustakas, “Countering Adversarial Attacks on Autonomous Vehicles Using Denoising Techniques: A Review,” *IEEE Open J. Intell. Transp. Syst.*, vol. 3, pp. 61–80, 2022, doi: 10.1109/OJITS.2022.3142612.

[141] A. K. Sahu and S. Kar, “Decentralized Zeroth-Order Constrained Stochastic Optimization Algorithms: Frank-Wolfe and Variants with Applications to Black-Box Adversarial Attacks,” *Proc. IEEE*, vol. 108, no. 11, pp. 1890–1905, 2020, doi: 10.1109/JPROC.2020.3012609.

[142] A. K. Nair, E. D. Raj, and J. Sahoo, “A robust analysis of adversarial attacks on federated learning environments,”

Comput. Stand. Interfaces, vol. 86, 2023, doi: 10.1016/j.csi.2023.103723.

[143] J. Chen, X. Lin, Z. Shi, and Y. Liu, "Link Prediction Adversarial Attack Via Iterative Gradient Attack," *IEEE Trans. Comput. Soc. Syst.*, vol. 7, no. 4, pp. 1081–1094, 2020, doi: 10.1109/TCSS.2020.3004059.

[144] S. Almutairi and A. Barnawi, "Securing DNN for smart vehicles: an overview of adversarial attacks, defenses, and frameworks," *J. Eng. Appl. Sci.*, vol. 70, no. 1, 2023, doi: 10.1186/s44147-023-00184-x.

[145] H. Xu *et al.*, "Adversarial Attacks and Defenses in Images, Graphs and Text: A Review," *Int. J. Autom. Comput.*, vol. 17, no. 2, pp. 151–178, 2020, doi: 10.1007/s11633-019-1211-x.

[146] Q. Li, C. Lin, Y. Yang, C. Shen, and L. Fang, "Adversarial Attacks and Defenses Against Deep Learning Under the Cloud-Edge-Terminal Scenes," *Jisuanji Yanjiu yu Fazhan/Computer Res. Dev.*, vol. 59, no. 10, pp. 2109–2129, 2022, doi: 10.7544/issn1000-1239.20220665.

[147] S. Kaviani, K. J. Han, and I. Sohn, "Adversarial attacks and defenses on AI in medical imaging informatics: A survey," *Expert Syst. Appl.*, vol. 198, 2022, doi: 10.1016/j.eswa.2022.116815.

[148] J. Li, Y. Liu, T. Chen, Z. Xiao, Z. Li, and J. Wang, "Adversarial attacks and defenses on cyber-physical systems: A survey," *IEEE Internet Things J.*, vol. 7, no. 6, pp. 5103–5115, 2020, doi: 10.1109/JIOT.2020.2975654.

[149] Z. Zhai, P. Li, and S. Feng, "State of the art on adversarial attacks and defenses in graphs," *Neural Comput. Appl.*, vol. 35, no. 26, pp. 18851–18872, 2023, doi: 10.1007/s00521-023-08839-9.

[150] S. A. Alsuhibany, "A Survey on Adversarial Perturbations and Attacks on CAPTCHAs," *Appl. Sci.*, vol. 13, no. 7, 2023, doi: 10.3390/app13074602.

[151] A. S. Albahri *et al.*, "A systematic review of trustworthy and explainable artificial intelligence in healthcare: Assessment of quality, bias risk, and data fusion," *Inf. Fusion*, vol. 96, pp. 156–191, 2023, doi: 10.1016/j.inffus.2023.03.008.

[152] A. S. Albahri *et al.*, "A Trustworthy and Explainable Framework for Benchmarking Hybrid Deep Learning Models Based on Chest X-Ray Analysis in CAD Systems," *Int. J. Inf. Technol. Decis. Mak.*, vol. 0, no. 0, pp. 1–54, 2024, doi: 10.1142/S0219622024500019.

[153] L. Alzubaidi *et al.*, "Towards Risk-Free Trustworthy Artificial Intelligence: Significance and Requirements," *Int. J. Intell. Syst.*, vol. 2023, p. 4459198, 2023, doi: 10.1155/2023/4459198.

[154] L. Alzubaidi *et al.*, "Comprehensive review of deep learning in orthopaedics: Applications, challenges, trustworthiness, and fusion," *Artif. Intell. Med.*, vol. 155, p. 102935, 2024, doi: https://doi.org/10.1016/j.artmed.2024.102935.

[155] M. A. Alsalem *et al.*, "Evaluation of trustworthy artificial intelligent healthcare applications using multi-criteria decision-making approach," *Expert Syst. Appl.*, vol. 246, p. 123066, 2024, doi: 10.1016/j.eswa.2023.123066.

[156] L. Alzubaidi *et al.*, "Trustworthy deep learning framework for the detection of abnormalities in X-ray shoulder images," *PLoS One*, vol. 19, no. 3 March, p. e0299545, 2024, doi: 10.1371/journal.pone.0299545.

[157] A. S. Albahri *et al.*, "A Systematic Review of Using Deep Learning Technology in the Steady-State Visually Evoked Potential-Based Brain-Computer Interface Applications: Current Trends and Future Trust Methodology," *Int. J. Telemed. Appl.*, vol. 2023, 2023, doi: 10.1155/2023/7741735.

[158] M. G. Yaseen and A. S. Albahri, "Mapping the Evolution of Intrusion Detection in Big Data: A Bibliometric Analysis," *Mesopotamian J. Big Data*, vol. 2023, pp. 138–148, 2023, doi: 10.58496/mjbd/2023/018.

[159] S. Rani, A. Kataria, S. Kumar, and P. Tiwari, "Federated learning for secure IoMT-applications in smart healthcare systems: A comprehensive review," *Knowledge-Based Syst.*, vol. 274, p. 110658, 2023, doi: 10.1016/j.knosys.2023.110658.

[160] M. R. Baker *et al.*, "Comparison of Machine Learning Approaches for Detecting COVID-19-Lockdown-Related Discussions During Recovery and Lockdown Periods," *J. Oper. Intell.*, vol. 1, no. 1, pp. 11–29, 2023, doi: 10.31181/jopi1120233.