STM JOURNALS

Journal of
Artificial Intelligence
Research & Advances

ISSN: 2395-6720
Volume 13, Issue 1, 2026
January–April
DOI Journal: 10.37591/JoAIRA

https://journals.stmjournals.com/joaira

Review JoAIRA

Adversarial Attacks on Machine Learning Models in Cybersecurity: A Systematic Literature Review

Sanjay Singh[1], Anamika Gupta[2,*]

Abstract

Adversarial machine learning (AML) is a field that is growing swiftly, especially as machine learning models are employed more and more in places where security is critical. This review goes into great depth over 746 publications from the Scopus database, with an emphasis on the connection between AML and network security. Using Biblioshiny and Scopus tools, we looked at trends in publications, study fields, productive authors, collaboration networks, and theme concentrations. Thirteen charts show how AML research has developed over time by highlighting notable contributions, main journals, popular keywords, and citation trends. Our research demonstrates that the topic is multidisciplinary and has research centers all around the world. It also shows that scholars are not working together as much anymore and that the focus is moving from pure attack modeling to robustness and interpretability. At the end of the study, some major gaps are pointed out, such as the clichés about real correspondent implementation and ethical governance. The report also offers approaches for future research to make the AML research ecosystem stronger and more accessible to everyone.

Keywords: Adversarial attack, adversarial machine learning, machine learning models, machine learning security, network security

INTRODUCTION

Biometric authentication and intrusion detection are two significant decision-making systems that currently incorporate machine learning models [1–3]. Their shortcomings are no longer merely technical details; they represent actual threats to real-world security. Adversarial machine learning (AML) is currently a major field of cybersecurity studies because it examines how inputs that are supposed to be detrimental might mislead or go around AI systems [4–8].

The convergence between AML and network security is crucial. In high-stakes circumstances, where a single error can lead to a phishing effort or data breach, it is vital to evaluate and limit adversarial risks [3, 9]. The field has achieved substantial advancements, such as determining where assaults can occur and how to protect against them [10–12]. However, research is still scattered across different topics, publication outlets, and geographic places [13–15].

This study uses bibliometric analysis to make sense of this difficult area of research. We examined 746 publications indexed in Scopus and built a structured map of the topic. This map depicts how the field has grown, what the main research areas are, who the most significant researchers are, how they cooperate, and where the most fascinating ideas are [16–18].

***Author for Correspondence**
Anamika Gupta
E-mail: anamikargupta@sscbsdu.ac.in

[1]PG Cybersecurity Student, Department of Computer Science, Shaheed Sukhdev College of Business Studies, University of Delhi, New Delhi, Delhi, India
[2]Professor, Department of Computer Science, Shaheed Sukhdev College of Business Studies, University of Delhi, New Delhi, Delhi, India

Received Date: November 03, 2025
Accepted Date: November 12, 2025
Published Date: December 11, 2026

Citation: Sanjay Singh, Anamika Gupta. Adversarial Attacks on Machine Learning Models in Cybersecurity: A Systematic Literature Review. Journal of Artificial Intelligence Research & Advances. 2026; 13(1): 23–38p.

This review is distinct from narrative reviews because it employs bibliometric tools such as preferred reporting items for systematic reviews and meta-analyses (PRISMA) and Biblioshiny to examine all the data. We do not simply want to summarize what we already know. We aim to uncover hidden trends, fill in the gaps, and provide guidance for future studies on AML, especially with regard to network security [19–24].

Questions for Research

RQ1: What kinds of assaults have been deployed against Machine Learning (ML) models in cybersecurity?

RQ2: What domains, datasets, and defenses are most typically studied?

RQ3: What patterns and gaps may be identified in the bibliometric analysis?

METHODOLOGY

Data Collection

The review began with a specialized search of the Scopus database using a mix of keywords, including "adversarial machine learning," "network security," "adversarial attack," and other relevant topics. The search was confined to English-language journal articles, conference proceedings, and reviews to make sure they were relevant and of good quality [25–28]. After removing entries that were not useful, the final set of 746 documents was combined (Figure 1).

Tools for Bibliometric Analysis

We used two key techniques to gain structured insights from the data:

- Biblioshiny (Bibliometrix R-package) generates performance measures including yearly publishing trends, top sources, authorship patterns, and theme mapping.
- Scopus analytics and human inspection: to back up what you see with more descriptive statistics and metadata.

These tools help us look at things on many levels, from big picture trends (such as the expansion of the field) to small picture dynamics (such as how productive each author is and how networks between institutions work).

Data Cleaning and Preprocessing

The dataset was cleaned for consistency before being displayed and analyzed. This included:

- Unifying author names (e.g., combining variations of the same author's name).
- Standardizing phrases and keywords, such as turning "adversarial ML" and "adversarial machine learning" into one word.

This procedure ensured that the results would be accurate, representative, and significant.

Visual Analytics Framework

A total of 13 visuals were generated, which served as the foundation for the analytical narrative built into the Results and Discussion sections, providing a bird's-eye view as well as deep dives into individual themes.

```
TITLE-ABS-KEY ((adversarial W/3 (attack+ OR example)) AND ("machine learning model*" OR "deep learning model*" OR "neural network*" OR "ML model*" OR classifier) AND (cybersecurity OR "network security" OR "information security")) AND NOT TITLE-ABS-KEY (image OR vision OR "object detection" OR "medical Imaging" OR "malware detection" OR phishing OR spam OR "intrusion detection" OR "anomaly detection" OR "botnet" OR "traffic classification" OR "domain detection" OR "wireless body area" OR "WBAN" OR "autonomous vehicle" OR "traffic sign")
```

Figure 1. An advanced query used for Scopus document search.

RESULTS

Publication Trend

Figure 2 shows the temporal pattern of publications (Documents by Year). Fewer than five papers were published annually between 2012 and 2016 on adversarial attacks in machine learning models for cybersecurity. This early stage demonstrates the field's infancy and restricted use of AML in security applications.

From 2018, a discernible shift was observed. Publications began to increase gradually, surpassing the 25-paper threshold in 2019. Between 2020 and 2023, growth is the steepest, reaching a peak of more than 175 documents by 2024 [14, 11, 17]. As ML deployment in cybersecurity matured, particularly with the rise of deep learning and neural networks in intrusion detection, spam filtering, and authentication systems, this surge has corresponded with a greater emphasis on adversarial robustness.

This timeline illustrates both the growing research community and the growing recognition of hostile threats as significant obstacles to reliable cybersecurity systems.

Geographic Distribution

Figure 3 highlights geographic research activity. China leads this category with more than 320 documents, representing approximately half of the collection. This considerable presence is unsurprising, considering China's significant spending on AI and cybersecurity research and its large academic network.

India is the third most active contributor, with approximately 80 documents, which implies a growing academic interest and consistency with worldwide research aims in AI security. Other countries, such as Australia, Canada, South Korea, and the United Kingdom, exhibit minor activities, each providing between 20 and 40 documents. France, Germany, and Hong Kong also figure in the top ten, although their contributions are fewer [29, 30].

This trend illustrates that while research is widespread, there is considerable concentration in Asia and North America. Collaboration among these countries could be an area for growth, potentially enhancing the generality of the findings and shared datasets.

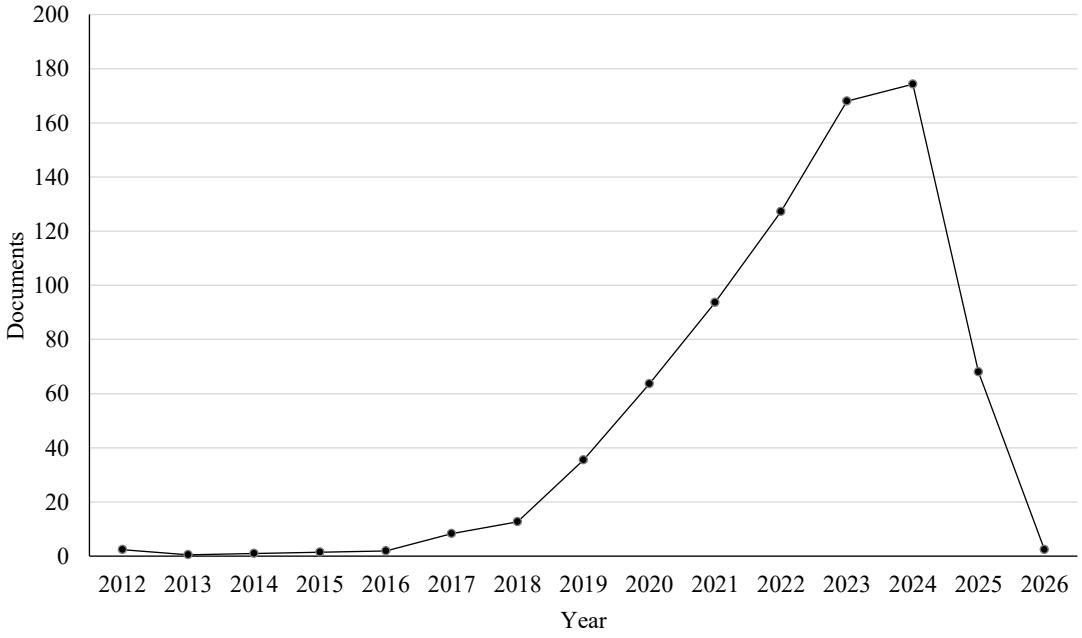

Figure 2. Documents by year.

Institutional Contributions

Figure 4 dives deeper into the specific institutional contributions. At the forefront is the Chinese Academy of Sciences, leading to approximately 30 papers, highlighting its critical role in advancing AML research within cybersecurity contexts.

Close behind are Tsinghua University and the Ministry of Education of the People's Republic of China, both with approximately 22–24 papers. These institutions are known for their interdisciplinary research and partnerships with national security and technology programs, which presumably explain their strong showings [13, 6, 8].

Outside Asia, the Pennsylvania State University appears on the list, indicating active participation from the United States in this field. These significant linkages underline that AML in cybersecurity is driven by a combination of academic and government research groups, many of which have a strategic interest in cyber defense.

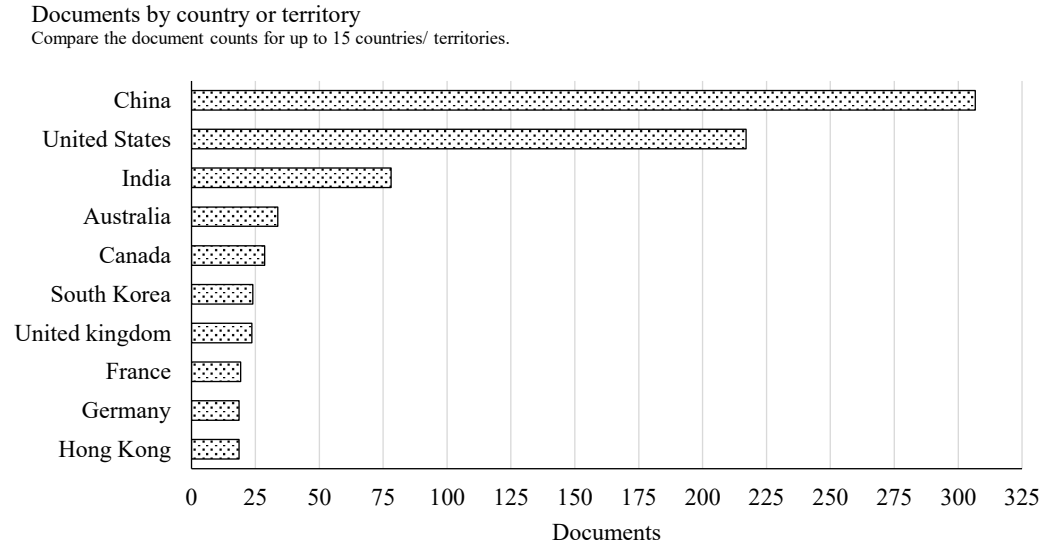

Figure 3. Documents by country or territory.

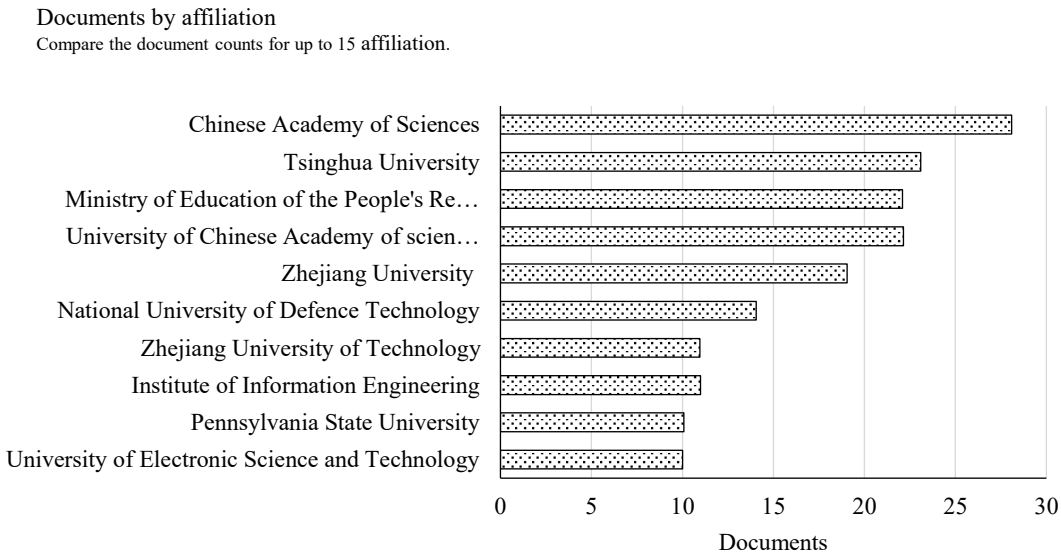

Figure 4. Documents by affiliation.

Subject Area Spread

Figure 5 provides insights into the disciplinary dispersion of literature. The majority of publications (45.4%) fall under computer science, which is expected, given that adversarial assaults and ML models are fundamental areas in AI research. The second greatest part, 22.1%, belongs to engineering, showing extensive work on applied systems, hardware-software integration, and practical implementation issues.

Mathematics (10.7%) and decision sciences (5.6%) imply that a large section of the area is also concerned with theoretical underpinnings, optimization approaches, and decision-making procedures under adversarial settings.

This subject area distribution demonstrates that while the topic is largely technical, it has started branching into cross-disciplinary work, an optimistic sign for future holistic approaches to cybersecurity [9, 16].

PRISMA Diagram

Figure 6 shows the PRISMA flow diagram detailing the study-selection process. A total of 746 records were identified in the Scopus database. No duplicates or ineligible records were removed at the identification stage, which indicates that the initial query was already well filtered.

All 746 records were included in the screening and eligibility assessments. No records were excluded after the assessment, which means that the inclusion criteria were broad enough to capture the intended literature without unnecessary filtering. The diagram provides transparency and replicability, which are the key requirements for the SLR methodology.

The lack of exclusions might suggest that although your query was broad, it remained focused on relevant topics after refinement, which is supported by the thematic alignment visible in other analyses [3, 7].

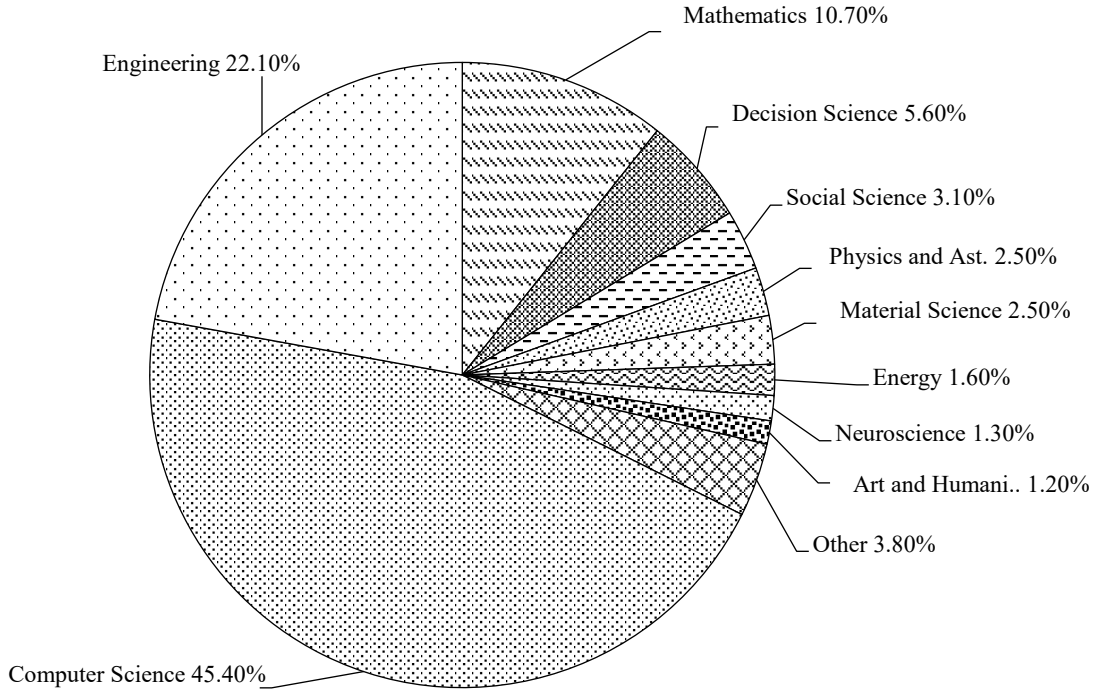

Figure 5. Documents by subject area.

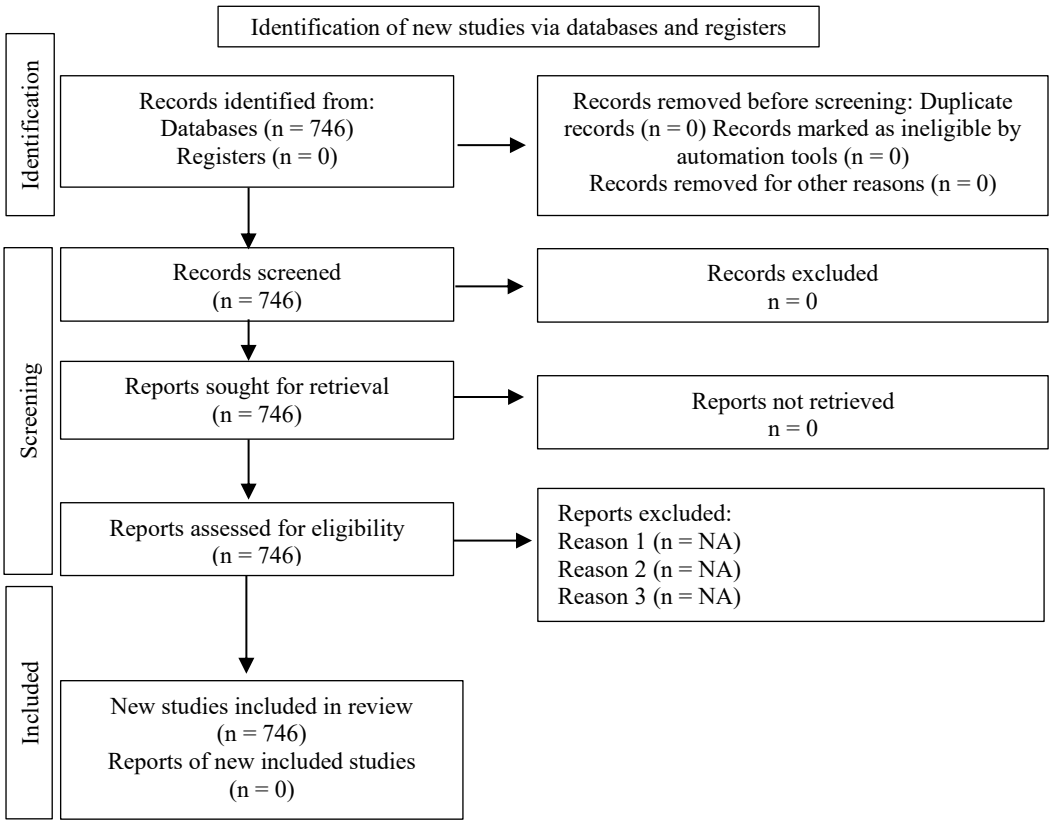

Figure 6. PRISMA flow diagram.

Annual Scientific Production

Figure 7 illustrates the annual growth in publications in the field. The annual scientific production plot shows the evolution of the research output over time. Between 2011 and 2016, the number of articles published each year remained negligible, indicating that this research area was still in its infancy or had not attracted widespread attention. From 2017 onward, there was a clear upward trend, with a modest rise between 2018 and 2019, followed by a steep acceleration from 2020. The publication count peaks sharply in 2023–2024, reaching over 170 articles, suggesting a surge in global interest and active research in this domain during that period. The noticeable drops in 2025 and 2026 are not necessarily indicative of a decline in research activity. This is more likely due to incomplete indexing for these years at the time of data collection, which is a common artifact in bibliometric studies. Overall, this trend reflects a rapidly growing field that has matured into a significant area of scholarly focus in recent years [20, 21].

Figure 8 maps the co-occurrence network of the authors' keywords to reveal thematic linkages in the literature. The keyword co-occurrence network highlights the intellectual structure of the research field by mapping how often specific keywords appear together in the analyzed literature. Two prominent clusters are identified.

The red cluster on the left is centered on the deep learning techniques. It includes terms such as deep neural networks, convolutional neural networks, adversarial defenses, attack methods, and perturbation techniques. These are closely interlinked, indicating that research on technical models and adversarial methods forms a dense thematic area.

On the right, the blue cluster groups broader security and machine learning terms such as machine learning, cybersecurity, AML, Internet of Things, and training algorithms. This cluster reflects the application-oriented side of the field, where machine learning concepts are integrated with security contexts.

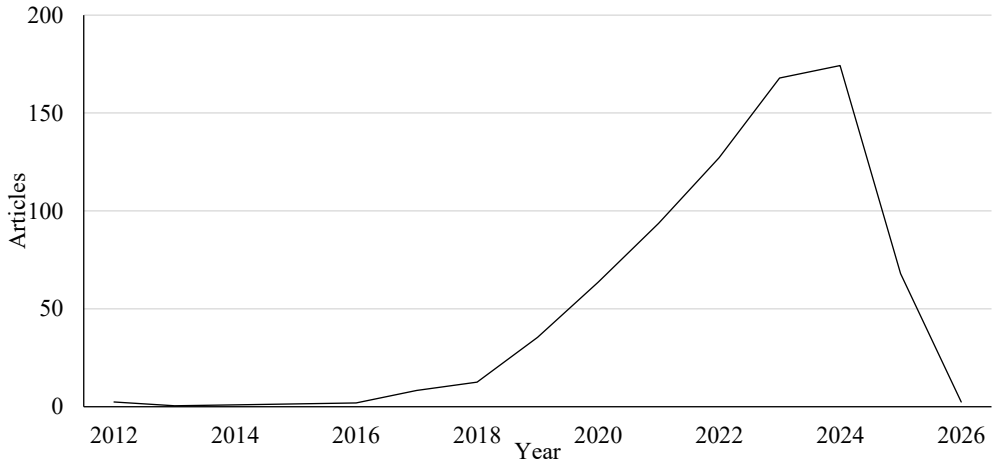

Figure 7. Annual scientific production.

Network security is the most frequently co-occurring keyword at the intersection of these clusters, and acts as a bridge between algorithmic research and practical security applications. This shows that the field is not isolated; instead, it sits at the intersection of advanced machine learning methods and their deployment in real-world security environments [1, 2, 31].

Authors' Keywords Co-Occurrence Network

Figure 8 maps the co-occurrence network of the authors' keywords to reveal thematic linkages in the literature. The keyword co-occurrence network highlights the intellectual structure of the research field by mapping how often specific keywords appear together in the analyzed literature. Two prominent clusters are identified.

The red cluster on the left is centered on the deep learning techniques. It includes terms such as deep neural networks, convolutional neural networks, adversarial defenses, attack methods, and perturbation techniques. These are closely interlinked, indicating that research on technical models and adversarial methods forms a dense thematic area.

On the right, the blue cluster groups broader security and machine learning terms such as machine learning, cybersecurity, AML, Internet of Things, and training algorithms. This cluster reflects the application-oriented side of the field, where machine learning concepts are integrated with security contexts.

Network security is the most frequently co-occurring keyword at the intersection of these clusters and acts as a bridge between algorithmic research and practical security applications. This shows that the field is not isolated; instead, it sits at the intersection of advanced machine learning methods and their deployment in real-world security environments [1, 2, 31].

Country Collaboration World Map

Figure 9 shows the international collaboration network through the country-to-country co-authorship links. The country collaboration map illustrates the geographical distribution of the research and the international partnerships driving it. The most prominent collaboration is between China and the United States, with 33 joint publications forming the backbone of global output in this field. Other strong links include China–Australia [8], China–Hong Kong [9], USA–Australia [9], and USA–Korea [14], indicating a high level of research exchange between the Asia-Pacific regions and North America.

Overall, this map demonstrates that research in this field is not confined to a single region. Instead, it is propelled by a global network of collaborations, with China and the United States acting as major hubs that connect researchers from diverse regions [32, 12].

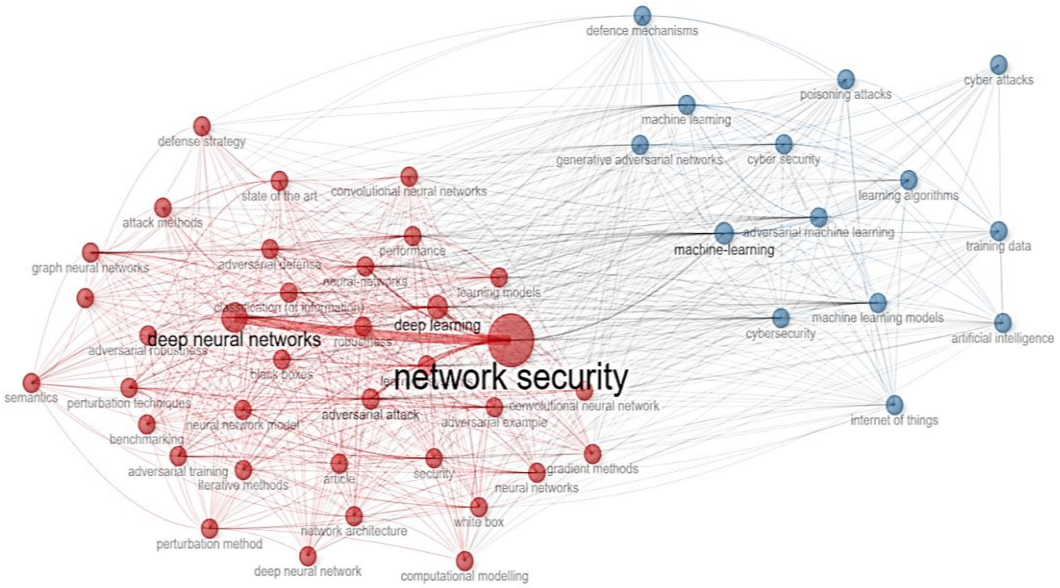

Figure 8. Authors' keywords co-occurrence network.

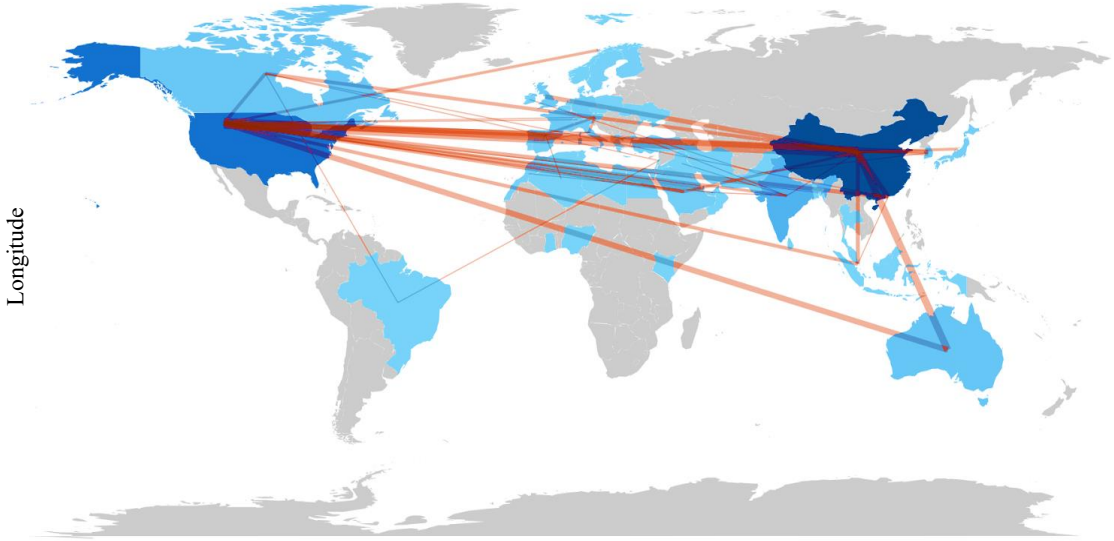

Figure 9. Country collaboration world map.

Factorial Analysis (Topics Dendrogram)

Figure 10 illustrates the factorial analysis dendrogram, indicating hierarchical groups of topics in the literature. The subject dendrogram obtained from factorial analysis presents a hierarchical perspective of how key terms cluster based on their co-occurrence patterns. Closely connected themes were grouped into branches to illustrate thematic ties.

For instance, one main branch ties perturbation approaches, robustness, and deep neural networks, showing a cluster focused on the model integrity and resilience.

Another branch links cybersecurity, artificial intelligence, AML, and machine learning models, establishing a theme area that concentrates on applying AI technologies to security concerns.

Separate clusters arise for the Internet of Things, semantics, and natural language processing systems, showing niche but increasing research subfields.

The structure of the dendrogram depicts how research subjects link and evolve, highlighting both well-established areas (such as adversarial assaults and defenses in deep learning) and peripheral themes (such as natural language processing (NLP) systems and internet of things (IoT) technologies) that are integrated into the broader debate. This visualization helps to map the intellectual landscape and highlight new linkages for future exploration [33, 34].

Word Cloud

Figure 11 presents an overview of the most commonly used keywords in the selected literature. The word cloud provides a visual depiction of author keywords that appear most often across the 746 papers. Keywords displayed in larger fonts were used more frequently, suggesting their significance in this field. Notably, "network security" appears to be the prevalent term, reinforcing the fundamental subject area of the reviewed literature.

Closely following are phrases such as "deep learning," "adversarial attack," "machine learning," and "adversarial machine learning"—all referring to the integration of artificial intelligence techniques in the context of cybersecurity. The prominence of "robustness", "perturbation", and "black box" implies that academics not only implement these models, but also critically assess their limitations and vulnerabilities to hostile inputs [11, 17, 18].

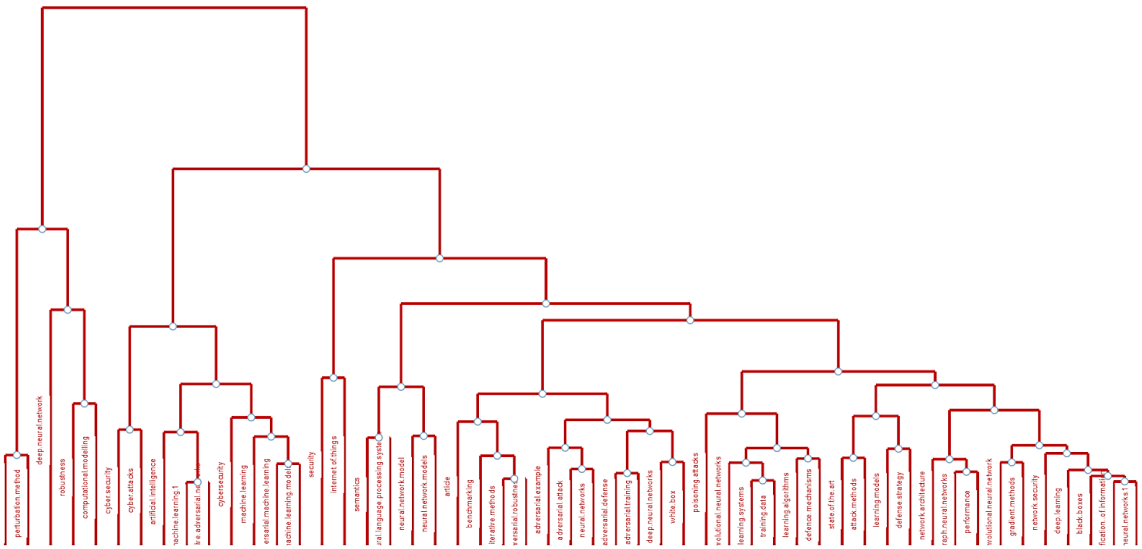

Figure 10. Factorial analysis.

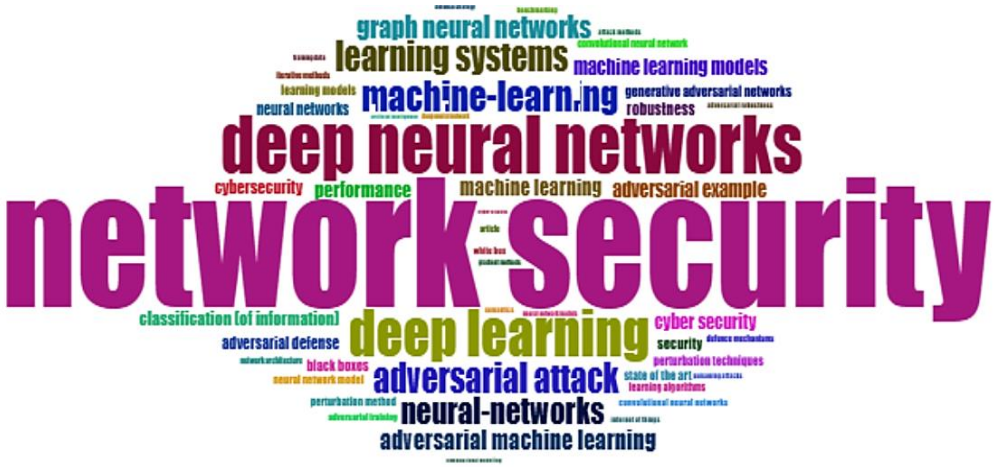

Figure 11. Word cloud.

Most Relevant Authors

Figure 12 illustrates the most relevant writers contributing to the topic field, based on publication frequency.

These graphics rank individual scholars according to the number of articles published in the collection. Chen Jinyin and Ji Shouling headed the list with seven publications, identifying them as important individuals in this scientific subject. Alouani Ihsen, Shen Chao, and Zheng Haibin follow closely, each providing six documents, demonstrating that they also have a considerable role in influencing the argument.

The generally balanced distribution of the top 20 authors reflects a collegial and dispersed research group. No single author dominates the field, alluding to a healthy range of perspectives and views. This also implies that the knowledge base is being developed by several separate groups, rather than focusing on a restricted circle.

Moreover, the participation of numerous authors from varied affiliations may imply inter-institutional or cross-national collaboration, which is crucial in sectors such as cybersecurity, where threat landscapes vary widely [5, 6, 35].

Most Relevant Sources

Figure 13 shows the most important publication sources based on the volume of documents contained.

The Lecture Notes in Computer Science (LNCS) series dominates this list with 43 documents, making it the most popular venue for publications in this dataset. This is expected given LNCS's wide coverage of international conferences on AI, security, and systems research.

Following the LNCS, IEEE Access accounts for 18 documents, providing an open-access platform that is popular for rapid dissemination. Computers and Security, a journal dedicated to the security domain, delivered 11 documents, demonstrating the necessity of expert channels for this subject.

This distribution also reflects the maturity of the domain—research is not limited to niche conferences but appears in well-established, high-impact publications [4, 10, 15].

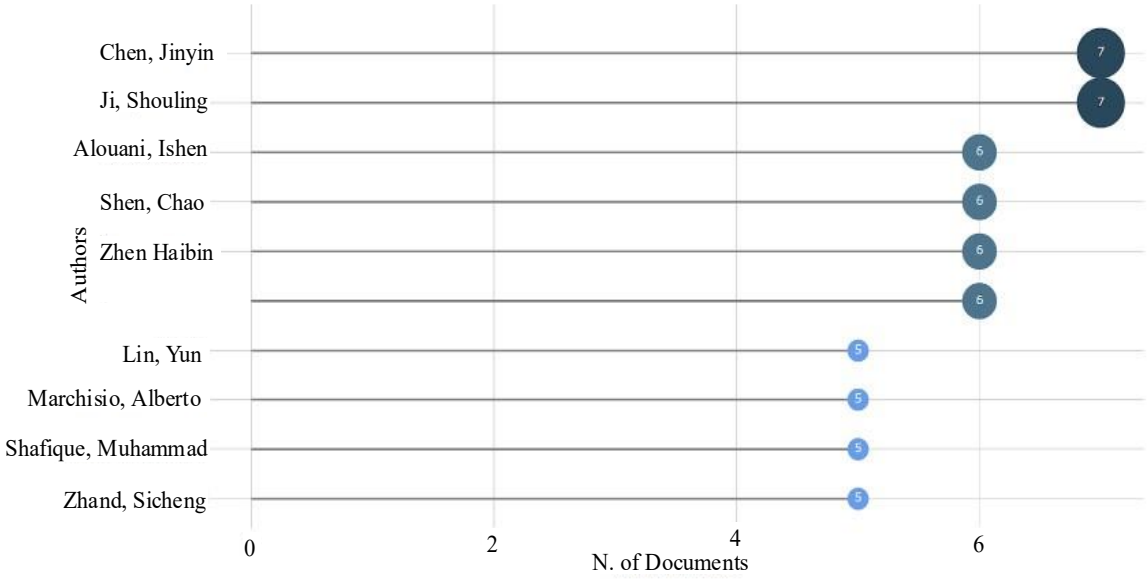

Figure 12. Most relevant authors.

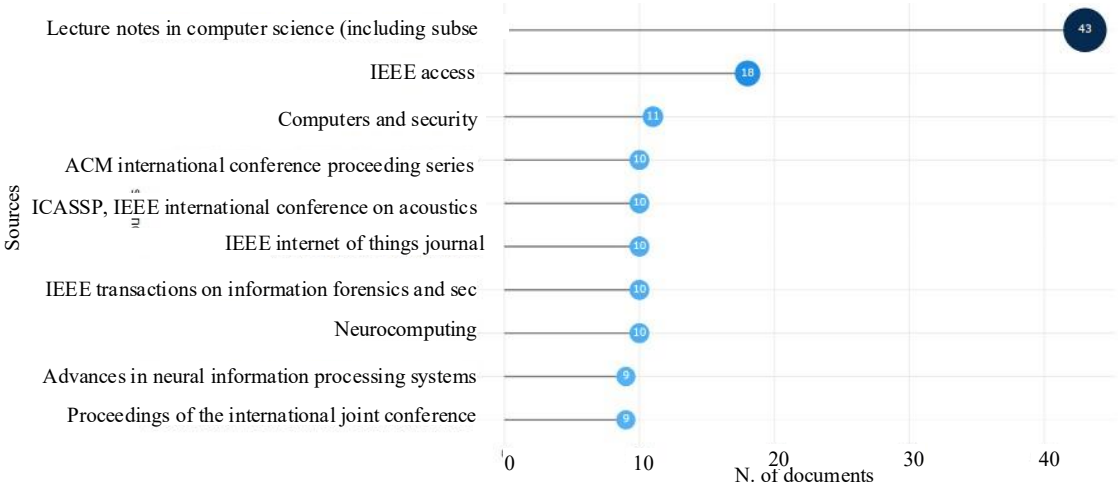

Figure 13. Most relevant sources.

General Overview of the Dataset

Figure 14 presents a general descriptive overview of the dataset used for the bibliometric analysis. Here is a summary that highlights the key components extracted from the core aspects of the analyzed studies.

Timespan covered: The dataset runs from 2012 to 2026. Although 2026 extends into the future, this presumably indicates early access to, or pre-published conference proceedings already indexed in the Scopus database.

Total publications: A total of 746 documents were analyzed and evaluated.

Author and collaboration metrics: The dataset comprises the contributions of 2,644 unique writers, with an average of 4.16 co-authors per document, indicating a high level of collaborative culture within the field. Notably, only 16 documents were single-authored, reflecting the tendency toward teamwork in this sector.

Keyword and citation analysis: Writers used 1,411 unique keywords, pointing to the thematic richness, diversity, and scope of the research. The papers collectively referenced 27,269 sources, with an average citation count of 26.61 per document, demonstrating that the field has acquired a reasonable academic impact [36, 37].

This bibliometric assessment and analysis illustrate the growing landscape of adversarial machine learning and its interactions with network security. Through an in-depth evaluation and examination of 746 papers acquired from the Scopus datasets, we deconstructed the disciplinary scope, publishing trends, important contributors, research hotspots, and collaboration tendencies. Together, the 13 visualizations provided a layered view of where the field sits and where it is heading.

The reproducibility of many suggested defenses has been questioned in a number of papers, including including recent works [28, 33], which have called for more standardized evaluation pipelines [25, 40]. The shortcomings of gradient-masking techniques were prominently revealed [41], highlighting the necessity of openness in experimental reporting. Additionally, the investigator [42] looked at threat models and pointed out discrepancies in attacker presumptions.

Furthermore, [37] promoted robust model distillation as a defense mechanism, whereas [38] concentrated on randomized smoothing as a certified defense. These methods offer viable substitutes for adversarial training.

Figure 14. General overview.

DISCUSSION

Thematic Trends and Research Maturity

The rapid expansion of research and publications over the years Figure 2 demonstrates that AML is a rapidly emerging research domain or subject, with research interest accelerating notably after 2018. This growth correlates with the increased adoption and integration of machine learning in security-sensitive domains, including authentication systems and malware detection technologies.

The disciplinary distribution, as shown in Figures 2 and 4, indicates that computer science and engineering remain dominant; however, there is a noticeable and growing contribution from mathematics and decision sciences, hinting at a more nuanced exploration of adversarial risks and defense strategies and mechanisms. Such multidisciplinarity is essential, especially when adversarial vulnerabilities can lead to tangible real-world impacts in high-stakes domains, such as healthcare, finance, and critical infrastructure.

Collaboration patterns by authors are shown in Figure 3, and Figure 13 further reveals a high co-authorship rate, with only 16 single-author articles out of 746 publications. This suggests that AML research is generally collaborative, typically requiring different expertise, from neural network architectures to cryptographic methods and threat modeling. Nevertheless, the relatively low number of international collaborations (~22.5%) implies that there is still untapped potential for cross-border cooperation, especially considering the global nature of cybersecurity challenges [24, 45, 27, 28, 46].

Influential Contributors and Communication Channels

Visualizations of the most cited documents from Figure 5, productive countries Figure 6, and top affiliations Figure 7 illustrate the primary geographical and institutional hotspots and organizational hubs of AML research. The United States, China, and India have led the way, sponsored by prominent institutions such as Tsinghua University and the Chinese Academy of Sciences. This geographic clustering has significant implications. While these countries dominate production, it may also reflect variations in their cybersecurity objectives, security goals, and threat models.

However, the citation network assessment in Figures 8 and 9 reveals that, while influential clusters exist, many scholars remain detached and isolated from each other's work. There is obvious fragmentation in which various groups work in parallel rather than in dialogue. These underlines needs for greater integrative evaluations and cross-group partnerships to knit together separate research initiatives within the AML research community [47, 48, 26, 28, 49–52].

Conceptual Focus and Research Gaps

Figures 10 and 13 help uncover the intellectual heartbeat of the field. The word cloud in Figure 10 highlights prominent subjects such as network security, deep learning, and adversarial assault. Simultaneously, the recurrence of terms such as robustness, explainability, and black box models speaks to a move from merely attacking adversarial threats towards developing interpretable and resilient machine learning systems. This demonstrates the healthy maturity of the domain, transitioning from identifying weaknesses to providing trustworthy solutions.

However, some research deficiencies stand out, such as:

- *Limited attention on real-world deployment:* While academic models abound, very few articles address deployment issues in production systems, especially in operational technology (OT), or integrate with environments where hostile defenses are difficult to install.
- *Lack of standard benchmarks:* The lack of universally approved evaluation metrics and benchmark datasets for adversarial defenses. This impairs reproducibility and comparative assessments.
- *Sparse work on regulation and policy:* Despite the significant ethical implications of hostile AI, particularly in areas such as surveillance, automated decision-making, and digital forensics, very few studies have focused on regulatory, legal, or societal issues.
- *Underexplored cross-dataset generalization:* Most models were evaluated using narrowly scoped datasets. In the future, researchers should strive toward robust generalization across varied input distributions and a wide range of threat scenarios.

Future Directions

Based on the findings and analysis, several promising directions emerge:

- *Bridging academic and industry silos:* More practitioner-oriented case studies and accessible benchmarks can significantly help reduce the gap between conceptual models and real-world deployment.
- *Interdisciplinary collaborations:* Integrating skills from behavioral science, cryptography, legal studies, and system engineering can boost adversarial resilience defense strategies.
- *Emphasis on explainability and ethics:* Future research must extend its focus beyond robustness to include explainability and accountability, especially in high-stakes domains such as healthcare and finance, where the consequences of AI decisions can be significant.
- *Increased international partnerships:* Promoting transnational collaboration can lead to more thorough threat modeling, especially when considering regional cyber dangers and legislation.

CONCLUSION

The changing field of AML in cybersecurity has been clarified by this bibliometric review. Growing awareness and urgency regarding adversarial threats have been indicated by the steady increase in research output since 2017. Collaborative research clusters addressing these threats from both theoretical and applied perspectives have been established by key contributors, including the USA, China, and India. The field's maturity is highlighted by recurring themes in various visualizations, including calls for explainability and robustness in real-world systems, emerging defense strategies, such as adversarial training and detection, and fundamental attack techniques, such as fast gradient sign method (FGSM) and projected gradient descent (PGD). Despite the progress, there are still gaps in the literature concerning adaptive threats, scalability, and model transferability in dynamic cyber environments.

Using bibliometric and thematic analyses, this review offers a cohesive perspective on the literature. The creation of international standards for adversarial robustness, integration with current cybersecurity frameworks, and defense generalization across domains should be top priorities for future research. Deploying safe, interpretable, and dependable AI systems in high-stakes situations will require such efforts. To ensure that actual robustness is achieved, it is also important to avoid obfuscation traps and use repeated benchmarks to measure the defense performance. It is important to improve threat model assumptions and add new information about certified defenses to the next generation of adversarial robust systems.

Acknowledgements

Sanjay Singh and Anamika Gupta would like to express their gratitude to the research communities and the open resources that enabled this review. This study was conducted independently using publicly accessible literature analyzed using Biblioshiny, Python, and other visualization tools. The literature was mainly accessed using Scopus.

The entire process, from conception to analysis and visualization, was performed independently and cooperatively. We value a larger research ecosystem that freely exchanges knowledge, allowing scholars to expand on previous research and return to the community.

REFERENCES

1. Babatunde LA, Etim ED, Essien IA, Cadet E, Ajayi JO, Erigha ED, et al. Adversarial machine learning in cybersecurity: Vulnerabilities and defense strategies. J Front Multidiscip Res. 2020;1:31–45. doi:10.54660/.JFMR.2020.1.2.31-45.
2. Eykholt K, Evtimov I, Fernandes E, Li B, Rahmati A, Xiao C, et al. Robust physical-world attacks on deep learning visual classification. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA. 2018. p. 1625–1634. doi:10.1109/CVPR.2018.00175.
3. Biggio B, Roli F. Wild patterns: Ten years after the rise of adversarial machine learning. In: Proc ACM SIGSAC Conf Comput Commun Secur. 2018. p. 2154–2156. doi:10.1145/3243734.3264418.
4. Akhtar N, Mian A. Threat of adversarial attacks on deep learning in computer vision: A survey. IEEE Access. 2018;6:14410–14430. doi:10.1109/ACCESS.2018.2807385.
5. Zhang J, Hu J, Liu J. Neural network with multiple connection weights. Pattern Recognit. 2020;107:107481. doi:10.1016/j.patcog.2020.107481.
6. Chen X, Liu C, Li B, Lu K, Song D. Targeted backdoor attacks on deep learning systems using data poisoning. [Preprint]. 2017. arXiv:1712.05526. doi:10.48550/arXiv.1712.05526.
7. Dilhara M, Ketkar A, Dig D. Understanding software-2.0: a study of machine learning library usage and evolution. ACM Trans Softw Eng Methodol. 2021 Oct;30(4):55. doi:10.1145/3453478.
8. Hendrycks D, Gimpel K. A baseline for detecting misclassified and out-of-distribution examples in neural networks. [Preprint]. 2016. arXiv:1610.02136. doi:10.48550/arXiv.1610.02136.
9. Dong Y, Liao F, Pang T, Su H, Zhu J, Hu X, Li J. Boosting adversarial attacks with momentum. In: Proc IEEE Conf Comput Vis Pattern Recognit (CVPR). 2018. p. 9185–9193. doi:10.1109/CVPR.2018.00957.
10. He J, Li X, Liao L, Song D, Cheung W. Inferring a personalized next point-of-interest recommendation model with latent behavior patterns. Proc AAAI Conf Artif Intell. 2016;30(1). doi:10.1609/aaai.v30i1.9994.
11. Gu T, Dolan-Gavitt B, Garg S. Badnets: Identifying vulnerabilities in the machine learning model supply chain. [Preprint]. 2017. arXiv:1708.06733. doi:10.48550/arXiv.1708.06733.
12. Liu Y, Chen X, Liu C, Song D. Delving into transferable adversarial examples and black-box attacks. [Preprint]. 2016. arXiv:1611.02770. doi:10.48550/arXiv.1611.02770.
13. Carlini N, Wagner D. Towards evaluating the robustness of neural networks. Proc IEEE Symp Secur Priv (SP). 2017;39–57. doi:10.1109/SP.2017.49.
14. Demontis A, Melis M, Pintor M, Jagielski M, Biggio B, Oprea A, et al. Why do adversarial attacks transfer? Explaining transferability of evasion and poisoning attacks. In: Proceedings of the 28th USENIX Security Symposium (USENIX Security 19). Berkeley (CA): USENIX Association; 2019. p. 321–338.
15. Wang X, Li J, Kuang X, Tan YA, Li J. The security of machine learning in an adversarial setting: A survey. J Parallel Distrib Comput. 2019;130:12–23. doi:10.1016/j.jpdc.2019.03.003
16. Ilyas A, Santurkar S, Tsipras D, Engstrom L, Tran B, Madry A. Adversarial examples are not bugs, they are features. In: Wallach HM, Larochelle H, Beygelzimer A, d'Alché-Buc F, Fox E, Garnett R, editors. Proceedings of the 33rd International Conference on Neural Information Processing Systems (NeurIPS 2019); 2019 Dec 8–14; Vancouver, BC, Canada. Red Hook (NY): Curran Associates Inc.; 2019. P. 125–136.
17. Jagielski M, Oprea A, Biggio B, Liu C, Nita-Rotaru C, Li B. Manipulating machine learning: Poisoning attacks and countermeasures for regression learning. 2018 IEEE Symposium on Security and Privacy (SP), San Francisco, CA, USA. 2018. p. 19–35. doi:10.1109/SP.2018.00057.
18. Li Y, Huang H, Guo X, Yuan Y. An empirical study on group fairness metrics of judicial data. IEEE Access. 2021;9:149043–149049. doi:10.1109/ACCESS.2021.3122443.
19. Madry A, Makelov A, Schmidt L, Tsipras D, Vladu A. Towards deep learning models resistant to adversarial attacks. [Preprint]. 2017. arXiv:1706.06083. doi:10.48550/arXiv.1706.06083.

20. Nasr M, Shokri R, Houmansadr A. Comprehensive privacy analysis of deep learning: Passive and active white-box inference attacks against centralized and federated learning. 2019 IEEE Symposium on Security and Privacy (SP), San Francisco, CA, USA. 2019. p. 739–753. doi:10.1109/SP.2019.00065.
21. Kurita K, Vyas N, Pareek A, Black AW, Tsvetkov Y. Measuring bias in contextualized word representations. In: Costa-jussà MR, Hardmeier C, Radford W, Webster K, editors. Proceedings of the First Workshop on Gender Bias in Natural Language Processing; 2019 Aug; Florence, Italy. Stroudsburg (PA): Association for Computational Linguistics; 2019. p. 166–172. doi:10.18653/v1/W19-3823.
22. Su J, Vargas DV, Sakurai K. One pixel attack for fooling deep neural networks. IEEE Trans Evol Comput. 2019;23:828–841. doi:10.1109/TEVC.2019.2890858.
23. Li J, Gao J, Jiang Q, He G. Adversarial defense networks via Gaussian noise and RBF. In: Sun X, Zhang X, Xia Z, Bertino E, editors. Artificial Intelligence and Security. ICAIS 2021. Lecture Notes in Computer Science. Vol. 12736. Cham: Springer; 2021. p. 480–491. doi:10.1007/978-3-030-78609-0_42.
24. Xiao C, Li B, Zhu JY, He W, Liu M, Song D. Generating adversarial examples with adversarial networks. In: Proceedings of the 27th International Joint Conference on Artificial Intelligence (IJCAI-18). Stockholm, Sweden: International Joint Conferences on Artificial Intelligence Organization; 2018. p. 3905-3911. doi:10.24963/ijcai.2018/543.
25. Sadeghi AR, Wachsmann C, Waidner M. Security and privacy challenges in industrial Internet of Things. In: Proceedings of the 52nd Annual Design Automation Conference (DAC '15); 2015 Jun; San Francisco, CA, USA. New York (NY): Association for Computing Machinery; 2015. Article 54, 6 p. doi:10.1145/2744769.2747942.
26. Cheng G, Sun X, Li K, Guo L, Han J. Perturbation-seeking generative adversarial networks: A defense framework for remote sensing image scene classification. IEEE Trans Geosci Remote Sens. 2022;60:1–11. doi:10.1109/TGRS.2021.3081421.
27. Alfakih T, Hassan MM, Gumaei A, Savaglio C, Fortino G. Task offloading and resource allocation for mobile edge computing by deep reinforcement learning based on SARSA. IEEE Access. 2020;8:54074–54084. doi:10.1109/ACCESS.2020.2981434.
28. Zhao ZQ, Zheng P, Xu ST, Wu X. Object detection with deep learning: A review. IEEE Trans Neural Netw Learn Syst. 2019;30:3212–3232. doi:10.1109/TNNLS.2018.2876865.
29. Li L. Comprehensive survey on adversarial examples in cybersecurity: Impacts, challenges, and mitigation strategies. [Preprint]. 2024. arXiv:2412.12217. doi:10.48550/arXiv.2412.12217.
30. Papernot N, McDaniel P, Jha S, Fredrikson M, Celik ZB, Swami A. The limitations of deep learning in adversarial settings. 2016 IEEE European Symposium on Security and Privacy (EuroS&P), Saarbruecken, Germany. 2016. p. 372–387. doi:10.1109/EuroSP.2016.36.
31. Jia R, Liang P. Adversarial examples for evaluating reading comprehension systems. In: Palmer M, Hwa R, Riedel S, editors. Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP 2017); 2017 Sep; Copenhagen, Denmark. Stroudsburg (PA): Association for Computational Linguistics; 2017. p. 2021–2031. doi:10.18653/v1/D17-1215.
32. Goodfellow IJ, Shlens J, Szegedy C. Explaining and harnessing adversarial examples. [Preprint]. 2014. arXiv:1412.6572. doi:10.48550/arXiv.1412.6572.
33. Finlayson SG, Bowers JD, Ito J, Zittrain JL, Beam AL, Kohane IS. Adversarial attacks on medical machine learning. Science. 2019;363(6433):1287–1289. doi:10.1126/science.aaw4399.
34. Pei K, Cao Y, Yang J, Jana S. DeepXplore: automated whitebox testing of deep learning systems. In: Proceedings of the 26th Symposium on Operating Systems Principles (SOSP '17); 2017 Oct; Shanghai, China. New York (NY): Association for Computing Machinery; 2017. p. 1–18. doi:10.1145/3132747.3132785.
35. Cortés-Pérez N, Torres-Méndez LA. A low-cost mirror-based active perception system for effective collision-free underwater robotic navigation. 2016 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Las Vegas, NV, USA. 2016. p. 61–68. doi:10.1109/CVPRW.2016.15.

36. Xie S, Girshick R, Dollár P, Tu Z, He K. Aggregated residual transformations for deep neural networks. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA. 2017. p. 5987–5995. doi:10.1109/CVPR.2017.634.
37. Nelson B, Barreno M, Chi FJ, Joseph AD, Rubinstein BIP, Saini U, et al. Exploiting machine learning to subvert your spam filter. In: Monrose F, editor. Proceedings of the First USENIX Workshop on Large-Scale Exploits and Emergent Threats (LEET '08); 2008 Apr 15; San Francisco, CA, USA. Berkeley (CA): USENIX Association; 2008.
38. Papernot N, McDaniel P, Sinha A, Wellman MP. SoK: Security and privacy in machine learning. 2018 IEEE European Symposium on Security and Privacy (EuroS&P), London, UK. 2018. p. 399–414. doi:10.1109/EuroSP.2018.00035.
39. Sharif M, Bhagavatula S, Bauer L, Reiter MK. Accessorize to a crime: Real and stealthy attacks on state-of-the-art face recognition. In: Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security; 2016; Vienna, Austria. New York (NY): ACM; 2016. p. 1528–1540. doi:10.1145/2976749.2978392.
40. Shokri R, Shmatikov V. Privacy-preserving deep learning. In: Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security; 2015; Denver, CO, USA. New York (NY): ACM; 2015. p. 1310–1321. doi:10.1145/2810103.2813687.
41. Song C, Ristenpart T, Shmatikov V. Machine learning models that remember too much. In: Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security; 2017; Dallas, TX, USA. New York (NY): ACM; 2017. p. 587–601. doi:10.1145/3133956.3134077.
42. Gan S, Zhang C, Qin X, Tu X, Li K, Pei Z, et al. CollaFL: Path sensitive fuzzing. 2018 IEEE Symposium on Security and Privacy (SP), San Francisco, CA, USA. 2018. p. 679–696. doi:10.1109/SP.2018.00040.
43. Tramèr F, Carlini N, Brendel W, Madry A. On adaptive attacks to adversarial example defenses. Adv Neural Inf Process Syst. 2020;33:1633–1645.
44. Szegedy C, Zaremba W, Sutskever I, Bruna J, Erhan D, Goodfellow I, et al. Intriguing properties of neural networks. [Preprint]. 2013. arXiv:1312.6199. doi:10.48550/arXiv.1312.6199.
45. Recht B, Roelofs R, Schmidt L, Shankar V. Do ImageNet classifiers generalize to ImageNet? In: Chaudhuri K, Salakhutdinov R, editors. Proceedings of the 36th International Conference on Machine Learning (ICML); 2019 Jun 9-15; Long Beach, CA, USA. Proc Mach Learn Res. 2019;97:5389-5400.
46. Zhang C, Bengio S, Hardt M, Recht B, Vinyals O. Understanding deep learning requires rethinking generalization. [Preprint]. 2016. arXiv:1611.03530. doi:10.48550/arXiv.1611.03530.
47. Zhang H, Chen H, Song Z, Boning D, Dhillon IS, Hsieh CJ. The limitations of adversarial training and the blind-spot attack. [Preprint]. 2019. arXiv:1901.04684. doi:10.48550/arXiv.1901.04684.
48. Xu H, Caramanis C, Mannor S. Robustness and regularization of support vector machines. J Mach Learn Res. 2009;10:1485-1510.
49. Ren K, Zheng T, Qin Z, Liu X. Adversarial attacks and defenses in deep learning. Engineering. 2020;6(3):346–360. doi:10.1016/j.eng.2019.12.012.
50. Ruiz N, Bargal SA, Sclaroff S. Disrupting deepfakes: adversarial attacks against conditional image translation networks and facial manipulation systems. [Preprint]. 2020 Mar 3. arXiv:2003.01279. doi:10.48550/arXiv.2003.01279.
51. Aria M, Cuccurullo C. Bibliometrix: An R-tool for comprehensive science mapping analysis. J Informetr. 2017;11:959–975. doi:10.1016/j.joi.2017.08.007.
52. Li N, Zhai L, Ma Z, Zhu X, Li Y. Lyapunov-guided deep reinforcement learning for service caching and task offloading in mobile edge computing. Comput Netw. 2024;250:110593. doi:10.1016/j.comnet.2024.110593.

