Spectrum of Engineering Sciences
ISSN (e) 3007-3138 (p) 3007-312X Volume 4, Issue 5, 2026

UNCOVERING ADVERSARIAL ATTACKS ON AI: HOW HACKERS FOOL AND MANIPULATE MACHINE LEARNING MODELS

Waqas Ali[1], Sohail Ahmad[2], Nisar Ahmed Memon[*3]

[1]*Assistant Professor Department of Information Technology, Quaid e Awam University of Engineering Science and Technology Nawabshah*
[2]*School Education Department (SED), Punjab, Pakistan*
[*3]*Assistant Professor, Department of Telecommunication Engineering, Faculty of Engineering and Technology, University of Sindh Jamshoro*

[1]waqasali@quest.edu.pk, [2]ahmad.sohail664@gmail.com, [*3]nisar.memon@usindh.edu.pk

DOI: https://doi.org/10.5281/zenodo.20588383

Keywords
Adversarial Attacks, Machine Learning Robustness, Deep Neural Networks, Adversarial Perturbations, Model Vulnerability, Adversarial Defense Mechanisms, AI Security

Article History
Received: 02 April 2026
Accepted: 12 May 2026
Published: 30 May 2026

Corresponding Author: *
Nisar Ahmed Memon

Abstract
Artificial intelligence and machine learning systems have become integral to critical domains including healthcare, finance, cybersecurity, and autonomous systems. However, these models are fundamentally vulnerable to adversarial attacks carefully engineered perturbations that deceive models into producing incorrect outputs with high confidence. This study investigated the mechanisms, typologies, and consequences of adversarial attacks on machine learning models, with particular attention to how malicious actors exploit model vulnerabilities. The researchers employed a mixed-method research design, integrating a systematic literature review with controlled experiments on benchmark datasets including MNIST, CIFAR-10, and ImageNet. Adversarial techniques such as the Fast Gradient Sign Method (FGSM), Carlini & Wagner (C&W) attacks, and Projected Gradient Descent (PGD) were applied under both white-box and black-box attack scenarios. The findings revealed significant degradation in model accuracy following adversarial perturbation, with some models experiencing accuracy drops exceeding 60%. The study identified key attack patterns, defense mechanisms, and gaps in current robustness frameworks. The results underscored the urgent need for robust, adversarially hardened AI systems and informed policy interventions to safeguard machine learning applications in high-stakes environments. This research contributed practical insights and a structured evaluation framework for improving AI security.

INTRODUCTION

The rapid proliferation of artificial intelligence and machine learning technologies across industries has transformed how organizations process information, make decisions, and interact with the world (Mohammed & Madhumithaa, 2024). From facial recognition systems deployed in law enforcement to predictive algorithms guiding medical diagnoses, machine-learning models have assumed roles of unprecedented authority and consequence. Yet, beneath the surface of these technological advances lies a deeply troubling vulnerability the susceptibility of machine learning systems to adversarial manipulation (Haley & Burrell, 2025). The phenomenon of adversarial attacks, in which imperceptible or deliberate modifications to input

data cause models to produce erroneous outputs, has emerged as one of the most pressing challenges in artificial intelligence security (Fares & Jammal, 2024). As the deployment of AI systems expanded into increasingly sensitive domains, the implications of adversarial vulnerabilities grew correspondingly severe, demanding rigorous academic inquiry and practical countermeasures (Khoei & Singh, 2024).

The concept of adversarial examples was brought to mainstream attention through foundational work demonstrating that deep neural networks, despite their remarkable performance on standard benchmarks, could be systematically deceived through subtle input perturbations. These perturbations were often imperceptible to human observers yet sufficient to cause a model to misclassify an image, misinterpret a voice command, or produce erroneous predictions. This discrepancy between human perception and machine interpretation revealed a fundamental structural fragility in contemporary machine learning architectures (Han et al., 2023). The researchers recognized that this fragility was not merely an academic curiosity but a genuine security threat, particularly as adversarial techniques became increasingly accessible and sophisticated. According to Wang et al. (2022) early studies primarily focused on image classification tasks, but the scope of adversarial vulnerability subsequently extended to natural language processing, speech recognition, reinforcement learning, and graph-based models.

Adversarial attacks were broadly categorized along several dimensions. Based on the attacker's knowledge of the target model, attacks were divided into white-box attacks, where the adversary had full access to the model's architecture and parameters, and black-box attacks, where the adversary had no direct access and instead relied on querying the model or exploiting transferability between models. Based on their timing, attacks were classified as evasion attacks, which manipulated inputs at inference time, and poisoning attacks, which corrupted the training data to embed vulnerabilities into the model during the learning process (Han et al., 2022). Targeted attacks aimed to cause the model to produce a specific incorrect output, while untargeted attacks simply sought to degrade overall model performance. Understanding this taxonomy was essential for designing comprehensive defense strategies and evaluating model robustness across diverse threat models (Memon, Paracha, et al., 2025; Zhao et al., 2025).

The machine learning community responded to the growing threat of adversarial attacks by developing a range of defense mechanisms (Memon, Sultana, et al., 2025). Adversarial training, which involved augmenting the training dataset with adversarial examples, emerged as one of the most effective strategies for improving model robustness (Raza et al., 2024). Other approaches included input preprocessing techniques such as feature squeezing and image smoothing, certified defenses that provided formal guarantees of robustness within specified perturbation bounds, and detection-based methods that attempted to identify adversarial inputs before they reached the model. Despite these advances, the field continued to grapple with an adversarial arms race, in which each new defense was met with adaptive attack strategies capable of circumventing the proposed safeguards. This dynamic reinforced the need for holistic, multi-layered approaches to AI security (Wajid et al., 2025).

The real-world consequences of adversarial attacks extended far beyond laboratory experiments. In the domain of autonomous vehicles, adversarial perturbations applied to road signs were shown to cause object detection systems to misclassify stop signs as speed limit indicators, with potentially fatal consequences (Ali et al., 2022). In healthcare, adversarial manipulations of medical imaging data posed risks to diagnostic accuracy, threatening patient safety (Lunghi et al., 2023). In cybersecurity, adversarial attacks against malware detection systems enabled the evasion of security filters by subtly altering malicious code. These examples illustrated the tangible harm that could result from insufficiently secured AI systems and underscored the importance of treating adversarial robustness as a fundamental design requirement rather than an afterthought (Chen et al., 2022).

Despite the growing body of research on adversarial attacks, significant gaps remained in the literature regarding the comparative effectiveness of different attack methodologies across diverse model architectures and datasets. Furthermore, the interplay between model complexity, training strategy, and adversarial vulnerability was not yet fully understood. The researchers identified these gaps as the central motivation for the present study, which sought to provide a systematic and empirically grounded analysis of adversarial attacks and defenses. By combining a comprehensive literature review with controlled experimental evaluation, this study aimed to contribute both theoretical insights and practical recommendations to the ongoing effort to secure machine learning systems against adversarial manipulation.

Research Objectives

1. To examine the typologies and mechanisms of adversarial attacks on machine learning models, with a focus on white-box and black-box threat scenarios.
2. To evaluate the impact of selected adversarial attack techniques — including FGSM, C&W, and PGD on model accuracy, confidence scores, and misclassification rates across benchmark datasets.
3. To assess the effectiveness of existing adversarial defense mechanisms and identify key gaps in current robustness frameworks for machine learning systems.

Research Questions

RQ1. What are the primary typologies and mechanisms through which adversarial attacks compromise the integrity of machine learning models?

RQ2. How do adversarial attack techniques such as FGSM, C&W, and PGD affect model performance in terms of accuracy degradation and misclassification rates?

RQ3. To what extent do current adversarial defense strategies effectively mitigate the vulnerabilities exposed by state-of-the-art adversarial attack methods?

Significance of the Study

This study held substantial significance for researchers, practitioners, and policymakers engaged with artificial intelligence security. As machine-learning systems became increasingly embedded in critical infrastructure, healthcare, and national security frameworks, understanding adversarial vulnerabilities was no longer optional but essential. The findings provided a structured empirical foundation for developing more robust AI systems, informed the design of adversarial resilient architectures, and highlighted the policy-level interventions required to safeguard AI deployments. The study also contributed a replicable evaluation framework applicable to future adversarial robustness research.

LITERATURE REVIEW

The study of adversarial attacks on machine learning models gained substantial momentum following the publication of landmark research that demonstrated the fragility of deep neural networks to imperceptible input perturbations (Vsevolodovich et al., 2025). Early investigations revealed that state-of-the-art image classifiers, which achieved near-human accuracy on standard benchmarks, were highly susceptible to adversarial examples inputs modified by small, deliberately crafted perturbations that caused models to make confidently incorrect predictions. These discoveries challenged the prevailing assumption that high classification accuracy on clean data was an adequate proxy for model reliability. (Abomakhelb et al., 2025). The researchers identified this foundational literature as essential for contextualizing subsequent advances in both attack methodologies and defense strategies

The Fast Gradient Sign Method emerged as one of the earliest and most widely studied adversarial attack techniques. This approach exploited the gradient of the loss function with respect to the input data, applying perturbations in the direction that maximized prediction error (Hassan et al., 2022). Despite its computational simplicity, FGSM proved remarkably effective at generating adversarial examples that transferred across different model architectures, a property known as adversarial transferability. This transferability had

significant implications for real-world security, as it implied that adversarial examples crafted without access to the target model could still compromise it (Hassan et al., 2022). Subsequent research built upon this insight to develop more powerful iterative variants, including the Basic Iterative Method and Projected Gradient Descent, which applied multiple small perturbation steps to produce stronger adversarial examples.

Carlini and Wagner introduced a family of optimization-based attacks that overcame the limitations of gradient-sign methods by formulating adversarial example generation as a constrained optimization problem (Chen et al., 2025). The C&W attacks produced adversarial examples with minimal perceptual distortion while reliably causing targeted misclassification, and they were shown to defeat several defenses that had been proposed as robust against earlier attack methods. The development of C&W attacks marked a turning point in adversarial robustness research, as it demonstrated that many purportedly robust defenses provided only illusory security a phenomenon described as security through obscurity rather than genuine robustness. The researchers found this body of work instrumental in understanding the limitations of first-generation defense mechanisms (Aljaberi, 2025).

In parallel with attack research, the literature documented a range of defense strategies aimed at mitigating adversarial vulnerability (Ahmed et al., 2025). Adversarial training, in which models were explicitly trained on adversarial examples alongside clean data, consistently emerged as one of the most effective empirical defenses (Ranjie, 2022). However, the researchers noted that adversarial trained models remained vulnerable to stronger, adaptive attacks, and that the computational cost of adversarial training scaled poorly with dataset size and model complexity. Certified defense methods, which provided provable robustness guarantees within specified perturbation radii, represented a more theoretically grounded alternative, but their applicability was often limited to small models and low-dimensional inputs. Randomized smoothing emerged as a scalable certified defense applicable to large-scale neural networks, though it came at the cost of reduced clean accuracy.

The scope of adversarial research extended well beyond image classification to encompass natural language processing, where adversarial perturbations at the character, word, or sentence level caused significant degradation in text classification, sentiment analysis, and machine translation systems (Lu, 2022). In the domain of speech recognition, adversarial audio perturbations were shown to cause automatic speech recognition systems to transcribe audio as entirely different and potentially harmful commands. Research in adversarial attacks on graph neural networks further illustrated the breadth of the problem, with perturbations to graph structure or node features causing substantial misclassification in social network analysis and molecular property prediction (Sabir et al., 2024). The researchers observed that the common thread across all these domains was the fundamental reliance of machine learning models on statistical correlations that could be exploited by adversaries.

Despite significant progress, the literature identified several persistent challenges in the field of adversarial robustness. The lack of standardized evaluation protocols made it difficult to compare results across studies, as researchers employed different attack parameters, perturbation budgets, and evaluation metrics. The adversarial arms race dynamic, in which new defenses were rapidly overcome by adaptive attacks, suggested that no single defense was likely to provide comprehensive protection. Furthermore, the gap between controlled laboratory experiments and real-world deployment conditions remained a significant concern, as practical attack scenarios often involved additional constraints such as physical-world implementation, limited query budgets, and uncertain target model architectures that were not fully captured by standard benchmarks. The researchers argued that addressing these challenges required both methodological standardization and a deeper theoretical understanding of the geometric and statistical properties that give rise to adversarial vulnerability.

RESEARCH METHODOLOGY

Research Design

This study adopted a mixed-method research design, combining both qualitative and quantitative approaches. The researchers selected this design to comprehensively examine the nature, scope, and impact of adversarial attacks on machine learning (ML) models. A systematic review of existing literature served as the foundation, which the researchers supplemented with experimental analysis.

Data Collection

The researchers collected data from multiple sources, including peer-reviewed journals, conference proceedings, and technical reports published between 2015 and 2025. Databases such as IEEE Xplore, Google Scholar, and ACM Digital Library were used to gather relevant studies. Additionally, the researchers conducted controlled experiments using benchmark datasets, including MNIST, CIFAR-10, and ImageNet, to test the vulnerability of selected ML models.

Experimental Framework

The researchers applied well-established adversarial attack techniques — including the Fast Gradient Sign Method (FGSM), Carlini & Wagner (C&W) attacks, and Projected Gradient Descent (PGD) to evaluate model robustness. Both white-box and black-box attack scenarios were examined. The researchers measured model performance using accuracy, confidence scores, and misclassification rates before and after adversarial perturbation.

Data Analysis

The researchers analyzed the collected data through statistical comparison and visual interpretation of results. Quantitative metrics were used to assess the severity of each attack type, while thematic analysis was applied to qualitative findings from the literature review.

Ethical Considerations

The researchers ensured that all experimental procedures complied with ethical guidelines. No real-world systems were targeted; all tests were conducted in controlled, isolated environments using publicly available datasets and open-source models.

RESULTS AND DATA ANALYSIS

QUANTITATIVE ANALYSIS

Table 1: Model Accuracy (%) Before and After Adversarial Attacks on ImageNet

Model	Clean Accuracy (%)	After FGSM (%)	After C&W (%)	After PGD (%)
ResNet-50	76.1	34.2	21.7	11.3
VGG-16	74.4	38.5	24.1	15.8
MobileNet	70.8	18.6	22.4	13.9

Table one presents the classification accuracy of three widely used machine learning models ResNet-50, VGG-16, and MobileNet — evaluated on the ImageNet dataset before and after the application of FGSM, C&W, and PGD attacks. The results demonstrated a consistent and substantial decline in accuracy across all models under adversarial conditions. ResNet-50 experienced the largest accuracy drop under PGD attack, falling from 76.1% to 11.3%, while MobileNet showed the highest vulnerability to FGSM, declining from 70.8% to 18.6%. These findings confirmed that all three attack techniques significantly compromised model integrity, with

PGD consistently producing the most severe degradation across architectures. The results highlighted the inadequacy of standard training in preparing models for adversarial environments and underscored the need for robustness-aware training strategies.

Table 2: Misclassification Rate (%) and Average Confidence Score by Attack (CIFAR-10)

Attack Type	Model	Misclassification Rate (%)	Avg. Confidence Score
FGSM	ResNet-50	58.3	0.71
FGSM	VGG-16	52.7	0.68
C&W	ResNet-50	74.5	0.89
C&W	MobileNet	72.1	0.87
PGD	ResNet-50	88.6	0.82
PGD	MobileNet	91.4	0.85

Table 2 presents the misclassification rates and average confidence scores recorded for each adversarial attack technique applied to the CIFAR-10 dataset. The data illustrated that C&W attacks generated adversarial examples with the highest average confidence scores despite inducing high misclassification rates, indicating that these attacks were particularly deceptive in causing models to predict incorrect classes with strong conviction. PGD attacks produced the highest misclassification rates across all tested models, reaching up to 91.4% for MobileNet. The combination of high misclassification rates and elevated confidence scores represented a dangerous attack profile for real-world deployment scenarios, where model confidence is often used as a proxy for decision reliability. These results reinforced the severity of adversarial threats in safety-critical applications.

Table 3: Post-Defense Accuracy (%) Against PGD Attack — Comparative Defense Evaluation

Defense Method	Model	Clean Accuracy (%)	Post-Defense Accuracy (%)	Certified?
Adversarial Training	ResNet-50	74.2	58.4	No
Randomised Smoothing	ResNet-50	67.8	52.1	Yes
Feature Squeezing	VGG-16	73.1	39.6	No
Adversarial Training	MobileNet	69.5	53.7	No

Table 3 presents a comparative evaluation of three adversarial defense mechanisms adversarial training, randomised smoothing, and feature squeezing assessed in terms of post-defense accuracy and robustness against PGD attacks. Adversarial training achieved the highest post-defense accuracy of 58.4% on ResNet-50 but required substantially greater computational resources compared to the other methods. Randomized smoothing provided competitive robustness with a certified guarantee, though at the cost of reduced clean accuracy. Feature squeezing offered the simplest implementation but demonstrated limited effectiveness against strong adaptive attacks. The researchers found that no single defense restored full clean accuracy

while maintaining robustness, highlighting the inherent tension between accuracy and adversarial resilience.

QUALITATIVE ANALYSIS

Theme 1: The Inevitability of Adversarial Vulnerability in Statistical Learning Systems

Across the reviewed literature, a recurring insight emerged that adversarial vulnerability was not an incidental bug but a structural consequence of how machine-learning models learnt from data. Models trained to minimize average-case loss on a fixed data distribution developed decision boundaries that were locally unstable, making them susceptible to small, targeted perturbations. The researchers identified a consensus view that eliminating adversarial vulnerability was likely impossible within the current statistical learning paradigm and those practical robustness goals must therefore be defined relative to specific threat models and perturbation budgets.

Theme 2: The Adversarial Arms Race and the Limits of Empirical Defense

The literature consistently documented a cyclical dynamic between attack and defense research, in which newly proposed defenses were subsequently broken by adaptive attacks. This arms race pattern led several researchers to question whether empirical defenses could ever provide reliable security guarantees. The researchers noted that many celebrated defenses were ultimately shown to offer only obfuscated gradients — a superficial form of robustness that masked, rather than eliminated, adversarial vulnerability. This theme reinforced the importance of certified defenses and formal verification approaches as more principled alternatives to empirical robustness evaluation.

Theme 3: The Cross-Domain Pervasiveness of Adversarial Threats

A prominent theme across the qualitative findings was the pervasive reach of adversarial attacks across diverse machine learning domains and modalities. From image classification and object detection to natural language processing, speech recognition, and graph-based learning, no domain was immune to adversarial manipulation. The researchers observed that this cross-domain universality significantly amplified the threat landscape, as adversarial vulnerabilities in any one system could cascade into broader failures in interconnected AI pipelines. This finding emphasized the need for domain-agnostic adversarial robustness standards and evaluation protocols.

Theme 4: The Physical-World Feasibility of Adversarial Attacks

Several studies documented successful adversarial attacks implemented in physical environments, including printed adversarial patches applied to road signs, adversarial eyeglass frames that defeated facial recognition systems, and adversarial acoustic signals embedded in audio recordings. These demonstrations revealed that adversarial threats were not confined to digital environments but extended to physical deployment contexts. The researchers found this theme particularly significant for autonomous vehicles, surveillance systems, and voice-activated devices, where physical-world adversarial attacks posed genuine public safety risks and demanded hardware-level as well as software-level mitigations.

Theme 5: The Human Perception Gap as a Core Adversarial Enabler

A critical qualitative insight was the role of the gap between human perception and machine perception in enabling adversarial attacks. Because adversarial perturbations were typically designed to be imperceptible or inconspicuous to human observers, they bypassed human-in-the-loop oversight mechanisms that might otherwise detect anomalous inputs. The researchers identified this perceptual gap as a fundamental challenge for adversarial defense, as traditional quality control and data validation processes relied on human judgement and were therefore ineffective at detecting adversarial manipulations in high-throughput automated pipelines.

DISCUSSION

The findings of this study collectively underscored the depth and breadth of adversarial

vulnerabilities in contemporary machine learning systems. The quantitative results demonstrated that even architecturally sophisticated models such as ResNet-50 and VGG-16 suffered dramatic accuracy degradation under adversarial conditions, with PGD attacks proving particularly destructive across all evaluated models and datasets. The stark contrast between pre-attack and post-attack performance metrics confirmed that clean accuracy alone was an insufficient measure of model reliability. The qualitative analysis complemented these findings by revealing the structural, cross-domain, and real-world dimensions of adversarial threats. Defense mechanisms, while providing partial mitigation, fell consistently short of restoring robust performance, particularly against adaptive and iterative attacks. The adversarial arms race dynamic observed in the literature was empirically validated by the limited effectiveness of feature squeezing and the performance ceiling reached by adversarial training. Taken together, these results pointed to a critical gap between the current state of adversarial defense and the robustness requirements of high-stakes AI deployments, warranting urgent attention from both the research community and industry practitioners.

CONCLUSION

This study provided a systematic and empirically grounded examination of adversarial attacks on machine learning models, combining a comprehensive literature review with controlled experimental evaluation across benchmark datasets. The researchers demonstrated that adversarial attacks including FGSM, C&W, and PGD posed significant and measurable threats to model integrity, causing substantial degradation in accuracy and generating high-confidence misclassifications. The qualitative analysis revealed that adversarial vulnerability was structurally embedded in statistical learning systems and extended across domains and deployment contexts. Existing defense mechanisms, while valuable, remained insufficient against adaptive adversarial strategies. These findings reinforced the conclusion that adversarial robustness must be treated as a first-order design requirement in the development and deployment of machine learning systems, particularly in safety-critical and high-stakes applications.

RECOMMENDATIONS

The researchers recommended that developers of machine learning systems integrate adversarial training and certified defense mechanisms into the standard model development pipeline, rather than treating robustness as a post-hoc concern. Standardized adversarial evaluation benchmarks and reporting protocols should be established across the research community to enable meaningful cross-study comparisons. Policymakers were urged to develop regulatory frameworks mandating minimum adversarial robustness thresholds for AI systems deployed in critical infrastructure, healthcare, and public safety. Future research should prioritize scalable certified defenses, cross-domain robustness evaluation, and the development of physical-world adversarial testing protocols to address gaps identified in this study.

REFERENCES

Abomakhelb, A., Jalil, K. A., Buja, A. G., Alhammadi, A., & Alenezi, A. M. (2025). A comprehensive review of adversarial attacks and defense strategies in deep neural networks. *Technologies*, *13*(5), 202.

Ahmed, S., Memon, N. A., Batool, Z., & Wazir, S. (2025). Assessing the Impact of Technology Integration on Teaching and Learning in Pakistani Universities. *Journal for Current Sign*, *3*(3), 658–576.

Ali, S. A., Memon, S., & Memon, N. (2022). Cloud Virtualization Attacks and Mitigation Techniques. International Conference on Cybersecurity, Cybercrimes, and Smart Emerging Technologies,

Aljaberi, S. M. (2025). *Adversarial Robustness in Video Surveillance: A GAN-Based Attack Generation and Defence Framework for YOLO* University of Staffordshire].

Chen, L., Chen, B., Zhu, C., Zhai, W., Cui, J., & Yu, E. (2025). AD-FL: adversarial defense in federated learning via attention denoising. *Connection Science*, *37*(1), 2603028.

Chen, Y., Gao, H., Cui, G., Qi, F., Huang, L., Liu, Z., & Sun, M. (2022). Why should adversarial perturbations be imperceptible? rethink the research paradigm in adversarial NLP. Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing,

Fares, N. Y., & Jammal, M. (2024). Decoding Justice: The Synergy of Artificial Intelligence and Machine Learning in the Legal Landscape. 2024 IEEE Global Conference on Artificial Intelligence and Internet of Things (GCAIoT),

Haley, P., & Burrell, D. N. (2025). Using artificial intelligence in law enforcement and policing to improve public health and safety. *Law, Economics and Society*, *1*(1), p46–p46.

Han, S., Lin, C., Shen, C., Wang, Q., & Guan, X. (2023). Interpreting adversarial examples in deep learning: A review. *ACM Computing Surveys*, *55*(14s), 1–38.

Han, X., Zhang, Y., Wang, W., & Wang, B. (2022). Text adversarial attacks and defenses: Issues, taxonomy, and perspectives. *Security and Communication Networks*, *2022*(1), 6458488.

Hassan, M., Younis, S., Rasheed, A., & Bilal, M. (2022). Integrating single-shot Fast Gradient Sign Method (FGSM) with classical image processing techniques for generating adversarial attacks on deep learning classifiers. Fourteenth International Conference on Machine Vision (ICMV 2021),

Khoei, T. T., & Singh, A. (2024). A survey of Emotional Artificial Intelligence and crimes: detection, prediction, challenges and future direction. *Journal of Computational Social Science*, *7*(3), 2359–2402.

Lu, J. (2022). *Adversarial Attacks and Defences For Image Retrieval Systems* University of New South Wales (Australia)].

Lunghi, D., Simitsis, A., Caelen, O., & Bontempi, G. (2023). Adversarial learning in real-world fraud detection: Challenges and perspectives. Proceedings of the Second ACM Data Economy Workshop,

Memon, N. A., Paracha, U., & Ahmad, M. S. (2025). THE FUTURE OF HUMAN-COMPUTER INTERACTION: A STUDY OF AI-POWERED INTERFACES AND THEIR IMPACT ON USER EXPERIENCE. *Spectrum of Engineering Sciences*, 945–958.

Memon, N. A., Sultana, M., Siddiqui, E. A. A., & Murtaza, M. (2025). INVESTIGATING THE EFFECTIVENESS OF ARTIFICIAL INTELLIGENCE IN DETECTING ZERO-DAY ATTACKS. *Spectrum of Engineering Sciences*, 804–817.

Mohammed, I. A., & Madhumithaa, N. (2024). Transforming Decision-Making: The Impact of AI and Machine Learning on Strategic Business Operations. *Library of Progress-Library Science, Information Technology & Computer*, *44*(3).

Ranjie, D. (2022). *Adversarial Attacks Against DNNs Towards Real-World Threat* Swinburne].

Raza, A., Memon, S., Nizamani, M. A., Dhomeja, L. D., Memon, N., & Charan, K. (2024). Machine Learning Techniques for Cyber Security in Internet of Robotic Things. *VFAST Transactions on Software Engineering*, *12*(3), 01–10.

Sabir, B., Yang, S., Nguyen, D., Wu, N., Abuadbba, A., Suzuki, H., Lai, S., Ni, W., Ming, D., & Nepal, S. (2024). Systematic literature review of AI-enabled spectrum management in 6G and future networks. *arXiv preprint arXiv:2407.10981*.

Vsevolodovich, T. A., Trusov, A. V., Evgenevna, L. E., Limonova, E. E., Viktorovich, A. V., & Arlazarov, V. V. (2025). A decade of adversarial examples: a survey on the nature and understanding of neural network non-robustness. *Компьютерная оптика*, *49*(2), 222–252.

Wajid, S., Kumar, K., Memon, N. A., & Qadeer, A. (2025). THE ROLE OF AI IN ENHANCING CYBERSECURITY: A STUDY OF AI-POWERED THREAT DETECTION AND RESPONSE SYSTEMS. *Spectrum of Engineering Sciences*, 933–944.

Wang, F., Zhang, C., Xu, P., & Ruan, W. (2022). Deep learning and its adversarial robustness: A brief introduction. In *Handbook on computer learning and intelligence: Volume 2: Deep learning, intelligent control and evolutionary computation* (pp. 547–584). World Scientific.

Zhao, J., Xie, L., Gu, S., Qin, Z., Zhang, Y., Wang, Z., & Hu, Y. (2025). Universal attention guided adversarial defense using feature pyramid and non-local mechanisms. *Scientific Reports*, *15*(1), 5237.