IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.67
Volume 5, Issue 7, March 2025

Adversarial Robustness of AI-Driven Claims Management Systems

Sita Rama Praveen Madugula and Nihar Malali
Independent researcher
praveenmsr@gmail.com and nihar.malali.r@gmail.com

Abstract: *Artificial intelligence (AI) has revolutionized claims management systems by streamlining processes such as fraud detection, document verification, and risk assessment, thereby enhancing operational efficiency and decision accuracy. However, AI-driven claims processing models are highly susceptible to adversarial attacks, where carefully crafted perturbations in input data can manipulate model predictions, leading to incorrect claim approvals, unjust denials, or exploitation by fraudulent actors. This study comprehensively investigates the adversarial robustness of AI-based claims management systems, analyzing different attack strategies, including evasion attacks that deceive models at inference time and poisoning attacks that corrupt training data to degrade model performance. Furthermore, it explore various defense mechanisms, such as adversarial training, robust feature extraction, uncertainty estimation, and model ensemble techniques, evaluating their effectiveness in mitigating vulnerabilities while balancing computational efficiency. Despite recent advancements, significant challenges persist in ensuring model robustness while maintaining accuracy, scalability, and compliance with evolving regulatory frameworks.*

Keywords: Claim Management Systems, Adversarial Robustness, Meta Learning, Blockchain Security

I. INTRODUCTION

AI has revolutionized claims management systems by automating processes such as fraud detection, document verification, and risk assessment[1]. These AI-driven systems leverage ML models to process vast amounts of structured and unstructured data, enabling faster and more accurate claims processing. However, despite their advantages, AI-powered claims management systems remain susceptible to adversarial attacks, where carefully crafted perturbations can manipulate model predictions, leading to incorrect claim approvals or denials[2]. Ensuring the robustness of these systems against adversarial threats is crucial for maintaining trust, security, and reliability in the financial and insurance sectors.

Adversarial attacks on claims management AI models typically fall into two categories: evasion and poisoning attacks. Evasion attacks involve modifying input data to deceive AI models into making incorrect classifications while poisoning attacks manipulate training data to degrade the system's performance over time[3]. These vulnerabilities can be exploited to commit fraud, bypass detection mechanisms, or generate incorrect risk assessments. Furthermore, the black-box nature of deep learning models makes it challenging to detect and mitigate such adversarial threats, necessitating advanced defense mechanisms.

Recent advancements in adversarial robustness have introduced techniques such as adversarial training, robust feature engineering, and model uncertainty estimation to improve the resilience of AI-driven claims management systems[4]. However, existing solutions often come at the cost of reduced model performance or increased computational overhead. Striking a balance between robustness and efficiency remains a critical challenge[5]. Additionally, regulatory compliance and ethical considerations further complicate the deployment of adversarial robust AI models in real-world claims management applications.

A. Motivation of the Study

AI-driven claims management systems have transformed the insurance and financial sectors by automating claim processing, fraud detection, and risk assessment. However, their susceptibility to adversarial attacks poses significant

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.67
Volume 5, Issue 7, March 2025

security and reliability concerns, as malicious actors can manipulate inputs to deceive AI models, leading to fraudulent approvals or unjustified rejections. Given the high-stakes nature of claims processing, where financial losses and customer trust are at risk, ensuring adversarial robustness is imperative. Existing defense mechanisms often trade off efficiency for security, highlighting the need for more resilient and scalable solutions. This study is motivated by the urgent need to enhance the adversarial robustness of AI-driven claims management systems, ensuring they can withstand sophisticated attacks while maintaining accuracy, fairness, and operational efficiency. The key contributions of the study are as follows:

- The paper explores the role of AI-driven automation in insurance claims processing, highlighting the benefits of machine learning and deep learning techniques for fraud detection, risk assessment, and workflow optimization.
- It categorizes different types of adversarial attacks, such as white-box, black-box, and poisoning attacks, and examines their impact on AI-driven claims management.
- The paper provides insights into how adversarial attacks affect various industries, including insurance fraud, healthcare diagnostics, cybersecurity, autonomous vehicles, and misinformation spread through deepfake manipulation.
- It reviews existing adversarial defense strategies, including adversarial training, defensive distillation, and model verification techniques, assessing their effectiveness in mitigating security risks in AI-powered claims processing.
- It posits that there is a need for better adversarial robustness in the context of insurance claims systems driven by AI, ethical AI frameworks and regulatory compliance to improve the reliability, transparency and security of AI-driven insurance claims systems.

B. Structure of the study

The study is structured as follows: In section I of the paper, AI in claims management is also introduced with the points where AI is used in automation and fraud detection. Section II talks about machine learning and deep learning models to be used for claims processing. In Section III adversarial threats are explored with white box and black box attacks. It undertakes an examination of defense mechanisms for improving AI robustness in Section IV. Section V concludes the study and gives directions for future research.

II. OVERVIEW OF AI IN CLAIMS MANAGEMENT SYSTEMS

In recent times, claims management is a cornerstone of improving operational efficiency in the insurance sector[6]. Process automation in insurance's economic implications such as the reduction of operational cost and the improvement of service quality through the reduction of human intervention[7]. It is based on integrating AI for claims automation, dedicated to real-time data processing and decision-making, which in turn has a direct impact on customer satisfaction.
There are many machine learning models for claim management with AI, including supervised learning for fraud detection, reinforcement learning for best claim management process workflow optimization and deep learning used for analyzing unstructured data[8]. AI usage by insurers alleviates the procedure of claims processing; decreases administrative intervention; and enhances accuracy in the process of risk assessment

A. Role of Machine Learning and Deep Learning Models

The ML and DL models are important for modern claims management systems. AI-based frameworks will help an organization to analyze and use big historical claims data to detect them, assess them and automate approval process. Presented A framework for automating, to a large extent, moderate automobile accident claims processing with the help of machine learning. Computer vision techniques for damage detection and repair cost estimation are introduced within the framework which allows insurers to decrease processing times[9]. The fact that AI-based tools intensified accuracy in assessing damages from property and vehicle claims helped in bringing the focus on more data-driven assessments.
Currently, CNNs are being utilized to perform deep learning to analyze image data in claims assessments in order to help insurers determine the severity of damage to vehicles and properties[10]. Transformer-based architectures are also used in natural language understanding for getting information from textual claims documents

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.67
Volume 5, Issue 7, March 2025

AI can help in faster, more accurate and with fewer errors processing of claims of insurance companies. AI can also deal with fraud and risk and enhance customer experience. Helps in automating the claims documentation. Through NLP, insurers can capture critical information from unstructured textual data, including claim descriptions, medical reports, and customer communications, with high accuracy.

That is, it showed how NLP algorithms are able to extract and categorize unstructured data from documentation with more than 80% accuracy in annotation tasks[11]. It lowers manual labor, decreases data analysis time and minimizes errors. It addressed the portability of NLP systems between healthcare institutions as that is an insurance claims documentation challenge and the adaptability and robustness of NLP tools.

Second, claims fraud detection, sentiment analysis and application of customer interaction automation have applied advanced NLP models such as BERT and LSTM networks[12]. Such models can look at large amounts of policyholder statements, social media interactions and legal documents and detect anomalies and avoid fraudulent claims.

B. Benefits and Risks of AI-Driven Automation in Insurance Claims

The acceptance of AI in the sphere of insurance claims management has its advantages, such as faster processing of claims, better efficiency that is to be expected, and a better customer experience. AI-driven automation decreases run costs, guarantees uniformity with regard to claim evaluations and decreases human errors[13].

Nevertheless, there are vast risks to AI-driven claims processing. Claims automation is vulnerable to adversarial attacks on AI models due to the fact that machine learning algorithms can be deceived by manipulated inputs and approve or deny the wrong claims[14]. Unfair claim assessments are also possible due to bias in training data and can harm policyholders. Further, there are data privacy concerns with AI-based claims systems as the insurers must comply with the regulatory framework such as GDPR and HIPAA to protect customer information[15].

Despite these challenges, insurers are increasingly investing in robust AI frameworks, integrating explainable AI (XAI) techniques to enhance transparency and regulatory compliance[16]. Future developments in data security, adversarial robustness, and ethical AI will be essential to guaranteeing the dependability of AI-powered claims management systems.

III. THREAT LANDSCAPE: ADVERSARIAL ATTACKS IN AI-BASED CLAIMS SYSTEMS

A. Types of Adversarial Attacks

Some common terms of adversarial attacks are presented.

Adversarial Example

A malicious example is a man-made example that introduces small changes to the original example in order to trick machine learning algorithms into making incorrect judgments while fooling human eyes into thinking the same thing[17].

White-Box Attack

Assuming complete control over the model's inputs and full knowledge of the model's structure (including its composition and partition layer settings), white-box attacks target models[18].

Black-Box Attack

Black-box attacks do not know how the model is internally structured; they can only manipulate the input and execute subsequent assaults by comparing the input and output feedback[19].

Real-World Attack/Physical Attack

Physical attacks in the real world don't know how the model is structured and even have limited input control[20].

Targeted Attack

In targeted assaults, the target is chosen before the attack, which leads to inaccurate predictions of the hostile pictures' unique labels. As a result, the impacts are assessed after the attacks[21].

Untargeted Attack

As long as the identification result is incorrect after the assault, untargeted attacks do not require setting the target beforehand[22].

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.67
Volume 5, Issue 7, March 2025

Evade Attack

Evade attacks include altering the input during the test phase and adding disruption to test samples in order to evade or fool the model's detection and prevent the AI model from being accurately recognized [23].

Poisoning Attack

Poisoning attacks create a backdoor or weakness in the learned model that may be exploited by attackers by including carefully crafted harmful instances during the model training process[24].

Backdoor Attack

A backdoor attack allows an attacker to access a system without requiring identity verification. This allows the attacker to get past security measures and cause more harm to the computer or network. In the neural network backdoor attack discussed in this paper, the attacker creates a model with a backdoor by inserting particular neurons into the neural network model. This allows the attacker to control the judgment on special inputs while maintaining consistency with the original model's judgment on normal inputs[25]. One kind of poisoning assault is a backdoor attack[26].

B. Real-World Implications of Adversarial Attacks

Adversarial attacks pose significant risks across various industries, affecting the reliability, security, and fairness of AI-driven systems. Their implications extend beyond theoretical vulnerabilities, impacting real-world applications in critical domains such as finance, healthcare, cybersecurity, and autonomous systems.

Financial Sector and Insurance Fraud

Adversarial attacks in AI-driven claims management systems allow for the manipulation of input data to avoid fraud detection mechanisms[27]. An example of this is that an attacker can subtly change claim documents or images of damaged property, making it seem as though the automated assessment models approve fraudulent claims. Likewise, such adversarial techniques can be used to fool risk evaluation models in banking, leading to incorrect granting or denying of loans.

Healthcare and Medical Diagnosis

AI models used for medical imaging and diagnostics are Adversarial perturbations that are also present in using AI for medical imaging and diagnostic tasks. If you modify an MRI scan that is so small and imperceptible to a human eye, you could cause a deep learning model to make a mistake and misclassify one kind of tumor as another kind; misdiagnose and mislead doctors into making harmful treatment decisions[28]. Therefore, ensuring the reliability of AI in healthcare requires the availability of adversarial robustness.

Cybersecurity and Malware Detection

It can be used further to bypass AI-based cybersecurity systems by way of adversarial attacks. For example, Malware samples may be modified by attackers to bypass the intrusion detection system (IDS) or the antivirus software based on machine learning[29]. The purpose is to permit this type of software to be used without detection from malicious software, resulting in data breaches, ransomware attacks and unauthorized access.

Autonomous Vehicles and Transportation

Adversarial attacks on computer vision systems in the field of autonomous driving can fool traffic signs or road marking[30]. For instance, a driverless car might potentially cause accidents by misinterpreting road signs, such as stop signs, as speed limits, leading to reckless driving. Ensuring adversarial robustness is essential to prevent accidents and ensure road safety.

Misinformation and Deepfake Manipulation

Adversarial attacks can also facilitate the spread of misinformation through manipulated content[31]. AI-generated deepfakes, enhanced through adversarial learning, can create realistic fake videos or audio clips that can be used for political manipulation, social engineering attacks, or reputational damage.

IV. ADVERSARIAL ROBUSTNESS IN AI-DRIVEN CLAIMS MANAGEMENT

The Adversarial defense mechanisms are techniques used to defend against attacks on machine learning models. These attacks can be launched by adversaries who try to limit or prevent accurate recognition performance. It is critical to invest in and immediately develop an unfavorably defensive model in order to withstand hostile attacks[32]. An example of a prominent work in the field of adversarial assaults on picture classification is DUNET, which successfully

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.67
Volume 5, Issue 7, March 2025

defended against both Whitebox and Blackbox attacks using a high-level representation-guided denoiser[33]. However, these protective procedures aren't well-suited to the environment of single-object monitoring. Though DUNET can only provide denoised images of a single, constant size, single-object tracking models necessitate the simultaneous input of template and search frames of varying sizes into the tracking model[34]. In addition, denoising the model does not protect it from adversarial assaults in tracking situations.

A. Adversarial Training

In order to make deep learning models more resistant to adversarial attacks, adversarial training is now thought to be the best practical approach. Nevertheless, adversarial training has a way to go before it can effectively deal with hostile threats[35]. Furthermore, deep learning models' generalizability is diminished as a result of adversarial training.

B. Defensive Distillation

Defensive distillation is based on training two copies of the same deep neural network architecture in sequence[36]. Readers are directed to the detailed presentation for more information about this, but to summaries, at high SoftMax temperatures, the model produces probabilities that are closer to a uniform distribution.

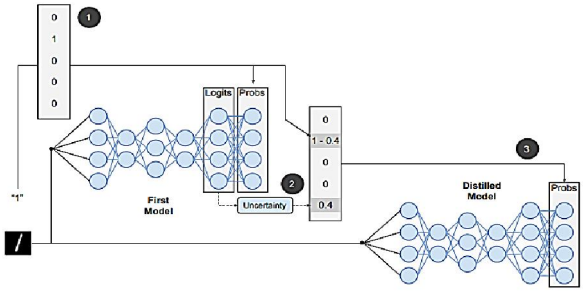

Figure 1: Defensive Distillation

Figure 1: The extended defensive distillation process consists of the following steps: (1) the first neural network is trained using one-hot labels as usual, (2) the original label information is combined with the predictive uncertainty of the first model, which is determined by making multiple stochastic passes through the model to infer the logits, and (3) a new labeling vector is defined for each training point. (3) Using the updated label vectors, the distilled model is trained at temperature $T \geq 1$.

C. Model Verification and Explainability Techniques

Explainable Methods for Artificial Intelligence. the types of XAI that can be applied in healthcare are briefly described in this section. There are numerous criteria used to categorize XAI techniques, according to recent literature. The classification criteria and associated categories for XAI methods are displayed in Figure 2.

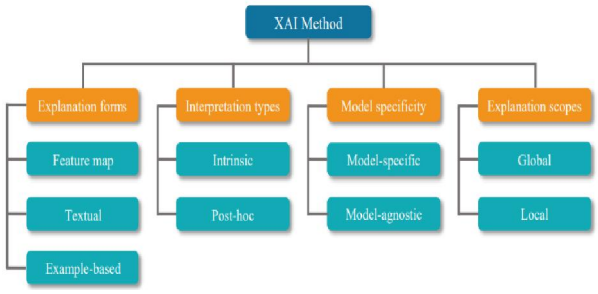

Figure 2: Categorization of explainable AI.

A taxonomy of XAI methods, categorizing them into four main dimensions: Figure 2 lists the types of explanations, their interpretations, model specificity, and explanation scopes. Explanation forms are feature maps, textual descriptions and example-based explanations[37]. Intrinsic (built-in explanations to the model) and Post hoc (explanation from after model output) are the ways of interpretation. There are between Model specific methods

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.67
Volume 5, Issue 7, March 2025

(specific for a given model) and Model agnostic methods (for any kind of model). Finally, explanation scopes categorize explanations as global and local (offering an overview of the model behavior as a whole or of an individual prediction, respectively)[38]. They put forth this structured approach to highlight the wide variety of tactics employed for model interpretability and transparency in XAI.

V. LITERATURE OF REVIEW

A literature review section the adversarial robustness of AI analyzes currently available techniques and methodologies in this section. Several methods are analyzed to improve the ability of AI systems to resist adversarial attacks.

Hyun and Park (2024) Through the design of additional self-generated auxiliary tasks within a realistic multi-task learning setup, research develops an efficient technique to increase adversarial robustness for main tasks. In order to accomplish effective multi-task learning without the need for predefined auxiliary tasks, their suggested method enables users to generate auxiliary tasks based on label information already present in their data. The newly created self-defined tasks play a supporting role in enhancing the primary task's adversarial accuracy while remaining "hidden" from attackers. Furthermore, the hidden auxiliary tasks make it possible to construct a rejection module that makes use of the auxiliary tasks' predictions in order to improve the accuracy of the prediction results[39].

Wei et al. (2024) approach involves linear variations, allowing for reversibility to recover the original images during the authorization process. They conduct experiments on various medical datasets, including different diseases and modalities. The results demonstrate significant decreases in medical image foundation models and standard models. These findings underscore that MIAD-MARK offers an effective, easily implemented, and robust solution to safeguard medical image copyright and patient privacy, thereby promoting the security of AI-driven medical image diagnosis in clinical applications[40].

Tuna and Kadan (2023) concentrate on adversarial assaults that target the distributed multiple-input multiple-output networks' power distribution. To lessen the impact of these attacks and contribute to improving the system's inherent performance, they suggest a unique defense strategy. The thorough simulations demonstrate that the suggested approach greatly improves the system's resilience. Adversarial attacks pose serious threats and have the ability to significantly undermine the security of AI-powered systems, particularly in sectors like communications where security is of utmost importance[41].

Nicolas et al. (2023) suggest a DNN model that makes use of certain parameters that have been adjusted to increase classification task accuracy. Model accuracy is a crucial performance indicator that may be enhanced by fine-tuning and architecture design. Three crucial loss functions—binary, categorical, and sparse categorical cross-entropy—are examined in order to achieve effective optimization. Using the sparse categorical cross-entropy loss function to reduce the error during training periods has shown excellent results. The study emphasizes how important accurate classification is and how powerful AI technologies might be used in power systems that support grid sustainability and stability[42].

Byun et al. (2022) utilize adversarial examples to their advantage. On the clean images, they find that these doubly adversarial examples occasionally veer towards different classes but generally return to the initial prediction. They use this information to suggest a regularization loss that reduces the risk of multi-targeted attacks by preventing these drifts. It is empirically demonstrated by experimental results on the CIFAR-10 and CIFAR-100 datasets that the suggested loss enhances adversarial robustness. Although adversarial examples can easily fool deep neural networks, they have demonstrated exceptional performance in a variety of domains [43].

Ramanathan and Jha (2021) illustrate the potential effects on the implementation of such experimental processes of AI systems' lack of adversarial resilience, such as protein folding networks. The establishment of a reliable and secure AI-driven AMP synthesis system may require addressing significant issues with adversarial robustness. A revolution in the planning, execution, and automation of scientific investigations is being driven by AI/ML tools. Particularly, expanding high-bandwidth equipment in conjunction with new hardware and software systems may greatly increase the throughput of experimental findings, and AI/ML approaches can offer previously unobtainable insights into cutting-edge research and ideas[44].

Table I provides the study based on the Adversarial Robustness of AI, including key objectives, focus, challenges and future work

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.67
Volume 5, Issue 7, March 2025

Table 1: Summary of literature review based on Adversarial robustness of AI

Study	Focus Area	Approach	Key Findings	Security/Privacy Concerns
Hyun and Park (2024)	Adversarial robustness in multi-task learning	Uses self-defined auxiliary tasks for improved robustness	Hidden auxiliary tasks enhance adversarial accuracy and enable a rejection module	Protects main tasks by hiding auxiliary tasks from attackers
Wei et al. (2024)	Medical image copyright and privacy protection	Uses linear variations for reversibility	MIAD-MARK ensures copyright protection and patient privacy in AI-driven medical image diagnosis	Safeguards medical images from unauthorized access
Tuna and Kadan (2023)	Adversarial attacks in power allocation for MIMO networks	Proposes a novel defense method to mitigate attacks	Increases robustness of power allocation in telecommunication networks	High security risks in telecom due to adversarial attacks
Nicolas et al. (2023)	Robust AI for power systems	Fine-tuned DNN model with optimized loss functions	Sparse categorical cross-entropy minimizes training error, enhancing AI robustness in power systems	Enhances AI-based power system security and grid stability
Byun et al. (2022)	Adversarial robustness using doubly adversarial examples	Proposes a regularization loss to prevent drift	Loss function strengthens the CIFAR-10 and CIFAR-100 datasets' resistance to multi-targeted assaults.	Decreases the susceptibility of deep neural networks to hostile instances
Ramanathan and Jha (2021)	Adversarial robustness in AI for protein folding networks	Highlights security gaps in AI-driven AMP synthesis	Identifies challenges in adversarial robustness for safe AI-driven scientific workflows	AI/ML-driven experiments require enhanced adversarial robustness

VI. CONCLUSION AND FUTURE WORK

AI-driven claims management systems have revolutionized the insurance industry by enhancing efficiency, reducing fraud, and automating decision-making. However, the dependability and security of these systems are seriously threatened by the growing complexity of hostile assaults. Numerous antagonistic threats, such as poisoning, black-box, and white-box assaults, have been investigated in this paper along with their practical ramifications. Furthermore, defense strategies such as robust feature engineering, adversarial training, and model uncertainty estimates were covered. Despite advancements in AI security, achieving a balance between robustness and efficiency remains a key challenge. Addressing adversarial vulnerabilities is crucial for maintaining trust and fairness in automated claims processing. The research has limitations because it lacks testing on authentic insurance database samples, which are representative of the full range of insurance scenarios. The research faces limitations in determining how well-proposed defense mechanisms function among various AI models.

Researchers should concentrate efforts on creating AI systems that remain attack-resistant, high-performing, and efficient. Hybrid defense systems that merge adversary training with explainable AI (XAI) would improve both their interpretation capabilities and their defensive strength. Federated learning techniques that secure sensitive data from poisoning assaults and blockchain integration for protected claims verification appear to be promising future strategic advancements. Additional studies must process regulatory and ethical factors to maintain compliance with current data protection regulations. The full security enhancement of AI-based claims management systems in practical applications will require joint action between academic institutions, researchers alongside industrial stakeholders and regulatory agencies.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.67
Volume 5, Issue 7, March 2025

REFERENCES

[1] S. Arora, S. R. Thota, and S. Gupta, "Artificial Intelligence-Driven Big Data Analytics for Business Intelligence in SaaS Products," in *2024 First International Conference on Pioneering Developments in Computer Science & Digital Technologies (IC2SDT)*, IEEE, Aug. 2024, pp. 164–169. doi: 10.1109/IC2SDT62152.2024.10696409.

[2] S. Arora, S. R. Thota, and S. Gupta, "Data Mining and Processing in the Age of Big Data and Artificial Intelligence - Issues, Privacy, and Ethical Considerations," in *2024 4th Asian Conference on Innovation in Technology (ASIANCON)*, IEEE, Aug. 2024, pp. 1–6. doi: 10.1109/ASIANCON62057.2024.10838087.

[3] S. A. and S. R. Thota, "Using Artificial Intelligence with Big Data Analytics for Targeted Marketing Campaigns," *Int. J. Adv. Res. Sci. Commun. Technol.*, vol. 4, no. 3, pp. 593–602, 2024, doi: DOI: 10.48175/IJARSCT-18967.

[4] S. Chatterjee, "Risk Management in Advanced Persistent Threats (APTs) for Critical Infrastructure in the Utility Industry," *Int. J. Multidiscip. Res.*, vol. 3, no. 4, pp. 1–10, 2021.

[5] S. Chatterjee, "Mitigating Supply Chain Malware Risks in Operational Technology : Challenges and Solutions for the Oil and Gas Industry," *J. Adv. Dev. Res.*, vol. 12, no. 2, pp. 1–12, 2021.

[6] K. S. Braunwarth, M. Kaiser, and A.-L. Müller, "Economic Evaluation and Optimization of the Degree of Automation in Insurance Processes," *Bus. Inf. Syst. Eng.*, 2010, doi: 10.1007/s12599-009-0088-6.

[7] J. Singh and S. Urolagin, "Use of Artificial Intelligence for Health Insurance Claims Automation," 2021. doi: 10.1007/978-981-15-5243-4_35.

[8] M. Gopalsamy and K. B. Dastageer, "The Role of Ethical Hacking and AI in Proactive Cyber Defense : Current Approaches and Future Perspectives," *Int. J. Innov. Sci. Res. Technol.*, vol. 10, no. 2, 2025, doi: https://doi.org/10.5281/zenodo.14916984.

[9] Y. Li, "Construction Project Claim Management under the Background of Wireless Communication and Artificial Intelligence," *Wirel. Commun. Mob. Comput.*, 2022, doi: 10.1155/2022/6074104.

[10] V. Pillai, "Anomaly Detection in Financial and Insurance Data-Systems," *J. AI-Assisted Sci. Discov.*, vol. 4, no. 2, 2024.

[11] D. Newman-Griffis *et al.*, "Linking Free Text Documentation of Functioning and Disability to the ICF With Natural Language Processing," *Front. Rehabil. Sci.*, 2021, doi: 10.3389/fresc.2021.742702.

[12] V. Kolluri, "A Pioneering Approach To Forensic Insights: Utilization AI for Cybersecurity Incident Investigations," *Int. J. Res. Anal. Rev.*, vol. 3, no. 3, 2016.

[13] V. Kolluri, "Cybersecurity Challenges in Telehealth Services : Addressing the security vulnerabilities and solutions in the expanding field of telehealth," *Int. J. Creat. Res. Thoughts*, vol. 8, no. 2, pp. 2348–1269, 2020.

[14] T. Li *et al.*, "Robust Image Recognition in Adversarial Environments : A Machine Learning Approach Robust Image Recognition in Adversarial Environments : A Machine Learning Approach," *researchgate.net*, 2023.

[15] S. R. Adavelli, "Autonomous Claims Processing : Building Self- Driving Workflows with Gen AI and ML in Guidewire," vol. 13, no. 12, pp. 1348–1357, 2024.

[16] S. P. Ramamoorthy, "AI in Pharmacy and Claims Management : Transforming Healthcare through Automation and Optimization AI in Pharmacy and Claims Management : Transforming Healthcare through Automation and Optimization Director of Automation at CVS Health , Texas , United Sta," no. November, 2024.

[17] N. Carlini, G. Katz, C. Barrett, and D. L. Dill, "Ground-Truth Adversarial Examples," *arXiv*, 2017.

[18] M. Nasr, R. Shokri, and A. Houmansadr, "Comprehensive privacy analysis of deep learning: Passive and active white-box inference attacks against centralized and federated learning," in *Proceedings - IEEE Symposium on Security and Privacy*, 2019. doi: 10.1109/SP.2019.00065.

[19] W. Xiao, H. Jiang, and S. Xia, "A New Black Box Attack Generating Adversarial Examples Based on Reinforcement Learning," in *2020 Information Communication Technologies Conference, ICTC 2020*, 2020. doi: 10.1109/ICTC49638.2020.9123270.

[20] Z. Zhou, B. Wang, M. Dong, and K. Ota, "Secure and Efficient Vehicle-to-Grid Energy Trading in Cyber Physical Systems: Integration of Blockchain and Edge Computing," *IEEE Trans. Syst. Man, Cybern. Syst.*, 2020, doi: 10.1109/TSMC.2019.2896323.

[21] Z. Katzir and Y. Elovici, "Why Blocking Targeted Adversarial Perturbations Impairs the Ability to Learn," *arXiv*. 2019.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.67
Volume 5, Issue 7, March 2025

[22] A. Wu, Y. Han, Q. Zhang, and X. Kuang, "Untargeted adversarial attack via expanding the semantic gap," in *Proceedings - IEEE International Conference on Multimedia and Expo*, 2019. doi: 10.1109/ICME.2019.00095.
[23] H. S. Anderson, A. Kharkar, B. Filar, D. Evans, and P. Roth, "Learning to Evade Static PE Machine Learning Malware Models via Reinforcement Learning," Jan. 2018.
[24] A. Demontis *et al.*, "Why do adversarial attacks transfer? Explaining transferability of evasion and poisoning attacks," in *Proceedings of the 28th USENIX Security Symposium*, 2019.
[25] B. Wang *et al.*, "Neural cleanse: Identifying and mitigating backdoor attacks in neural networks," in *Proceedings - IEEE Symposium on Security and Privacy*, 2019. doi: 10.1109/SP.2019.00031.
[26] T. D. Nguyen, P. Rieger, M. Miettinen, and A.-R. Sadeghi, "Poisoning Attacks on Federated Learning-based IoT Intrusion Detection System," 2021. doi: 10.14722/diss.2020.23003.
[27] A. S. Liu, J. Guo, S. M. Li, Y. S. Xiao, X. L. Liu, and D. C. Tao, "A Survey on Adversarial Attacks and Defenses for Deep Reinforcement Learning," *Jisuanji Xuebao/Chinese J. Comput.*, 2023, doi: 10.11897/SP.J.1016.2023.01553.
[28] A. K. MI Khan, A Arif, "AI-Driven Threat Detection: A Brief Overview of AI Techniques in Cybersecurity," *BIN Bull. Informatics*, vol. 2, no. 2, pp. 248–61, 2024.
[29] S. Nokhwal, S. Nokhwal, S. Pahune, and A. Chaudhary, "Quantum Generative Adversarial Networks: Bridging Classical and Quantum Realms," in *2024 8th International Conference on Intelligent Systems, Metaheuristics & Swarm Intelligence (ISMSI)*, New York, NY, USA, NY, USA: ACM, Apr. 2024, pp. 105–109. doi: 10.1145/3665065.3665082.
[30] S. Regmi and S. Regmi, "Threats of Adversarial Attacks on Deep Learning," *Int. J. Res. Appl. Sci. Eng. Technol.*, vol. 11, no. 9, pp. 885–891, Sep. 2023, doi: 10.22214/ijraset.2023.55716.
[31] S. B. Shah, "Machine Learning for Cyber Threat Detection and Prevention in Critical Infrastructure," *J. Glob. Res. Electron. Commun.*, vol. 2, no. 2, pp. 1–7, 2025, doi: https://jgrec.info/index.php/jgrec.
[32] A. K. Aftab Arif, Muhammad Ismaeel Khan, "An overview of cyber threats generated by AI," *Int. J. Multidiscip. Sci. Arts*, vol. 3, no. 4, pp. 67–76, 2024.
[33] W. Suttapak, J. Zhang, H. Zhao, and L. Zhang, "Multi-Model UNet: An Adversarial Defense Mechanism for Robust Visual Tracking," *Neural Process. Lett.*, vol. 56, no. 2, pp. 1–19, 2024, doi: 10.1007/s11063-024-11592-2.
[34] A. J. Rahul Dattangire, Ruchika Vaidya, Divya Biradar, "Exploring the Tangible Impact of Artificial Intelligence and Machine Learning: Bridging the Gap between Hype and Reality," *2024 1st Int. Conf. Adv. Comput. Emerg. Technol.*, pp. 1–6, 2024.
[35] T. Bai, J. Luo, J. Zhao, B. Wen, and Q. Wang, "Recent Advances in Adversarial Training for Adversarial Robustness," *IJCAI Int. Jt. Conf. Artif. Intell.*, no. 2, pp. 4312–4321, 2021, doi: 10.24963/ijcai.2021/591.
[36] N. Papernot and P. McDaniel, "Extending Defensive Distillation," May 2017.
[37] M. M. Srinivas Murri, Manoj Bhoyar, Guru Prasad Selvarajan, "Transforming Decision-Making with Big Data Analytics: Advanced Approaches to Real-Time Insights, Predictive Modeling, and Scalable Data Integration," *Int. J. Commun. Networks Inf. Secur.*, vol. 16, no. 5, pp. 506–519., 2024.
[38] S. Tyagi, "Analyzing Machine Learning Models for Credit Scoring with Explainable AI and Optimizing Investment Decisions," *Am. Int. J. Bus. Manag.*, vol. 5, no. 01, pp. 5–19, 2022.
[39] C. Hyun and H. Park, "Multi-Task Learning with Self-Defined Tasks for Adversarial Robustness of Deep Networks," *IEEE Access*, 2024, doi: 10.1109/ACCESS.2024.3355924.
[40] X. Wei, B. Pu, C. Chen, and H. Fu, "MIAD-MARK: Adversarial Watermarking of Medical Image for Protecting Copyright and Privacy," in *2024 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, 2024, pp. 3842–3845. doi: 10.1109/BIBM62325.2024.10822032.
[41] O. F. Tuna and F. E. Kadan, "A Novel Method to Mitigate Adversarial Attacks on AI-Driven Power Allocation in D-MIMO," in *2023 IEEE International Black Sea Conference on Communications and Networking, BlackSeaCom 2023*, 2023. doi: 10.1109/BlackSeaCom58138.2023.10299750.
[42] D. Nicolas, H. Figueroa, Y. Wang, W. Elmannai, and G. C. Giakos, "Adversarial Machine Learning Architecture in AI-Driven Power Systems," in *2023 IEEE International Conference on Dependable, Autonomic and Secure Computing, International Conference on Pervasive Intelligence and Computing, International Conference on Cloud and Big Data Computing, International Conference on Cyber Science and Tec*, 2023. doi: 10.1109/DASC/PiCom/CBDCom/Cy59711.2023.10361400.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.67
Volume 5, Issue 7, March 2025

[43] J. Byun, H. Go, S. Cho, and C. Kim, "Exploiting Doubly Adversarial Examples for Improving Adversarial Robustness," in *2022 IEEE International Conference on Image Processing (ICIP)*, IEEE, Oct. 2022, pp. 1331–1335. doi: 10.1109/ICIP46576.2022.9897374.

[44] A. Ramanathan and S. K. Jha, "Adversarial Attacks against AI-driven Experimental Peptide Design Workflows," in *Proceedings of XLOOP 2021: 3rd Annual Workshop on Extreme-Scale Experiment-in-the-Loop Computing, Held in conjunction with SC 2021: The International Conference for High Performance Computing, Networking, Storage and Analysis*, 2021. doi: 10.1109/XLOOP54565.2021.00010.