STM JOURNALS

Journal of
Operating Systems
Development & Trends

ISSN: 2454-9355
Volume 13, Issue 1, 2026
January–April
DOI (Journal): 10.37591/JoOSDT

https://journals.stmjournals.com/joosdt

Review JoOSDT

Navigating the Dual Edge: A Comprehensive Technical Survey of Security, Privacy, and Countermeasures in the Era of Artificial Intelligence

Shubhashree Pattanayak[1,*], Suman Sahoo[2], Sanjay Kumar Sahoo[3]

Abstract

Artificial intelligence (AI) is seamlessly woven into vital sectors, such as self-driving cars, high-speed trading systems, and defense strategies; it has triggered a counterintuitive development in advanced cyberattacks. This survey paper attempts to perform an in-depth technical analysis of the "AI attack surface." There are threats across three main vectors. Data integrity attacks focus specifically on examining "clean-label" poisoning and backdoor injection. Model confidentiality breaches discuss the mathematics behind model inversion and membership inference attacks. Generative exploitation examines prompt injection and hallucinations in large models. This paper presents a comprehensive technical survey of the dual nature of AI: its potential to both bolster security and be exploited as an attack vector. We first catalog and analyze the principal threat models associated with AI systems, including adversarial examples, model inversion, data poisoning, and inference attacks. We then explore privacy vulnerabilities stemming from training data leakage, unauthorized model access, and collaborative learning paradigms like federated learning. Through this lens, the survey examines how traditional threats evolve in AI contexts and highlights new attack vectors unique to learning-based systems. Next, we systematically review state-of-the-art countermeasures across defensive categories—robust training, certified defenses, differential privacy, cryptographic approaches, and secure multi-party computation—emphasizing both strengths and limitations. Fast gradient sign method (FGSM) and projected gradient descent (PGD) are examples. In addition, we investigate the combined realms of AI and the Internet of Things (IoT) using the ZIRCON framework. Instead, we promote a need for a "zero-watermarking" approach. In closing, we have a strategic outlook for 2025–2075. In fact, we believe that "explainability" or XAI is not just a norm for regulation. Rather, it is a key to unlocking a future full of artificial general intelligence.

Keywords: Adversarial machine learning, data poisoning, differential privacy, Explainable AI (XAI), Internet of Things (IoT) security, large language models, prompt injection, ZIRCON

***Author for Correspondence**
Shubhashree Pattanayak
E-mail: pattnayakshubhashree555@gmail.com

[1]Assistant Professor, Department of Computer Science and Engineering, Gandhi Institute of Excellent Technocrats, Bhubaneswar, Odisha, India
[2,3]Student, Department of Computer Science and Engineering, Gandhi Institute of Excellent Technocrats, Bhubaneswar, Odisha, India

Received Date: January 27, 2026
Accepted Date: January 31, 2026
Published Date: March 20, 2026

Citation: Shubhashree Pattanayak, Suman Sahoo, Sanjay Kumar Sahoo. Navigating the Dual Edge: A Comprehensive Technical Survey of Security, Privacy, and Countermeasures in the Era of Artificial Intelligence. Journal of Operating Systems Development & Trends. 2026; 13(1): 1–6p.

INTRODUCTION

Artificial intelligence (AI) has matured from being a theory-based field to a presence factor, paradigmatically transforming the spectrum of healthcare, finance, and military operations. However, this omnipresence comes at a high cost. As rightly stated by Paracha et al. [1], "AI is a 'dual edge' weapon; it is a force multiplier for military forces since it provides threat detection capabilities with automation on a massive scale, but it is also now being leveraged for automating complex attacks on a massive scale."

The importance of this dichotomy is further exemplified by the metrics that have been developed. According to the Stanford AI Index Report (2025), there has been a "56.4% year-over-year increase in AI-specific security incidents." Unlike more conventional vulnerabilities that are regularly found in standard software programming and often follow logical errors in the code (such as those behind 'buffer-overflow' attacks), those that appear in AI models are often stochastic and reflect the internal nature of the learning process [2]. It is possible to have a mathematically sound model that 'converges perfectly' during the learning process and is nevertheless 'statistically vulnerable to adversarial attack in high-dimensional spaces.'

The classic view of cybersecurity was perimeter security, which includes firewalls, intrusion detection, and access controls. However, with the advent of the age of AI, the "perimeter" no longer exists; model parameters now include sensitive data, with inputs (images) acting as attack vectors. This shift is historically evident: while the 1990s–2010s focused on attacking infrastructure (Distributed denial of service (DDoS) attacks), the period from 2015 to 2023 saw a transition towards model evasion (adversarial examples) [3]. From 2024 to the present, attacks have evolved to include semantic vectors, such as prompt injections and hallucination induction, as well as supply chain vulnerabilities, such as model serialization attacks. In this study, we sought to compile an exhaustive classification of such potential risks. For our analysis, we focus on breaking it down by the cycle of AI: training phase (poisoning), inference phase (evasion/privacy), and deployment phase (generative risks and edge security) [4].

LITERATURE REVIEW AND RELATED WORK

The field of adversarial machine learning has expanded rapidly in the past decade. Early research has demonstrated that machine learning models, particularly deep neural networks, can be misled by carefully crafted inputs containing subtle and nearly imperceptible perturbations, commonly referred to as adversarial examples. These minor modifications, although invisible to human observers, can cause models to produce highly confident but incorrect predictions. Subsequent work introduced efficient gradient-based techniques for generating such adversarial inputs, showing that vulnerability arises largely from linear behavior in high-dimensional feature spaces rather than solely from overfitting or excessive model complexity [5].

In addition to adversarial examples, privacy-related threats have become a major concern. It has been demonstrated that attackers can determine whether a specific data sample is part of a model's training dataset, a technique commonly known as a membership inference attack. These risks have accelerated the adoption of differential privacy principles in machine learning to ensure that individual training records cannot be reverse-engineered from model outputs [2, 1].

Recently, the rapid rise of large language models has introduced new security challenges, including prompt manipulation and injection attacks. Security organizations have formally recognized these risks and identified critical vulnerabilities in data pipelines, model architecture, and deployment infrastructures. The concept of an "AI attack surface" encompasses all potential exposure points within these systems [6].

Data poisoning attacks target the integrity of training datasets by injecting malicious or manipulated samples to influence the model's behavior. Availability attacks aim to degrade the overall model performance, sometimes leading to a denial-of-service effect by overwhelming the system with excessive or low-quality data that prevents proper convergence. Integrity attacks are more subtle because they attempt to induce specific, targeted misclassifications while maintaining a high overall model accuracy [7].

A particularly sophisticated form of poisoning is a clean-label attack. Unlike traditional attacks that rely on incorrect labeling, this method preserves the correct labels while subtly altering the input

features [8]. For example, an attacker may embed a small, nearly invisible trigger pattern into images of a specific class. During training, the model learns to associate the hidden trigger with the class label. After deployment, the same trigger can be applied to different objects, causing the model to misclassify them with high confidence. Because the labels remain correct and the modifications are difficult to detect visually, such attacks can bypass standard human verification processes and remain undetected in the training data [9].

DISCUSSION AND ANALYSIS

Adversarial evasion occurs after a model is trained and deployed, exploiting the gradient information of the neural network to find "pockets" in the input space where the model fails. One of the foundational techniques in this domain is the fast gradient sign method (FGSM). The FGSM is a "one-shot" attack where θ represents the model parameters, x the input, y the true label, and $J(\theta, x, y)$ the loss function. The adversary seeks a perturbation η such that $\|\eta\|\infty \leq \varepsilon$ (where ε is sufficiently small to be invisible) that maximizes the loss. The FGSM calculates this by taking the sign of the gradient as follows:

$$\eta = \varepsilon \cdot \text{sign}(\nabla_x J(\vartheta, x, y)) \quad (1)$$

The adversarial example is then generated as:

$$x_{ady} = x + \eta \quad (2)$$

Intuitively, this formula calculates the direction of the steepest ascent of the loss function, relative to the input data. By taking a small step (ε) in this direction, the attack accumulates a massive change in the output activation, exploiting the high-dimensional linearity of the model [10].

Although the FGSM is fast, it is not the most robust attack. The projected gradient descent (PGD) attack is an iterative version of the FGSM. It applies the gradient step multiple times, projecting the result back onto the ε-ball around the original input after each step.

$$x^{t+1} = \Pi_{x+S}(x^t + \alpha \cdot \text{sign}(\nabla_x J(\theta, x^t, y))) \quad (3)$$

This equation represents an iterative optimization. At each step t, the input is adjusted by a step size α in the gradient direction. The projection operator Π then ensures that the new input remains within a defined valid range (the ε-ball) of the original image, ensuring that the perturbation remains imperceptible.

In terms of real-world implications [5], the presented case studies involve Open Pilot, an open-source autonomous driving system. Tests showed that adversarial patches placed on roads could trick the system into swerving into the oncoming traffic. These "Physical World" attacks are harder to execute than digital ones because of lighting and angle variations, but they represent a lethal threat to cyber-physical systems [11].

Deep learning models also act as information sponges, optimizing loss by memorizing patterns that often include specific training examples. This leads to privacy leakage risks, such as inference attacks. Model Inversion attempts to reconstruct the input data x from the model output f (x). For example, in a facial recognition system, if an attacker knows the name of a Person A (e.g., "Person A"), they can optimize a random image to maximize the confidence score for that class:

$$x* = \text{argmax}_x \quad P\ (y = \text{Person A}|x) \quad (4)$$

This optimization problem seeks to find the input $x*$ that yields the highest probability for a specific target label y; by applying gradient ascent, a random initial image slowly morphs into a ghostly but recognizable version of the victim's face, effectively reconstructing the private data.

Another threat is the membership inference attack (MIA), which is a binary classification problem

that determines whether a specific record x was used to train model M. This relies on the observation that models behave differently on training data (lower loss, higher confidence) than on the unseen data. Attackers train "Shadow Models" to mimic the target and learn the statistical signature of membership.

To counter these threats, differential privacy (DP) provides a mathematically rigorous definition of privacy. A randomized algorithm M is (ε, δ)-differentially private if, for all adjacent datasets D and D′ (differing by one element) and for all subsets of outputs $S \subseteq \text{Range}(M)$:

$$P(M(D) \in S) \leq e^{\varepsilon} \cdot P(M(D') \in S) + \delta \quad (5)$$

Here, ε (epsilon) represents the privacy budget. A lower ε tightens the bound, meaning that the outputs of the model are more indistinguishable regardless of whether a specific individual's data are in the dataset, thus guaranteeing higher privacy [12].

The most common mechanism to achieve this is DP-SGD (Differentially Private Stochastic Gradient Descent), which involves clipping gradients to limit individual influence and adding noise.

$$g_{noisy} = \text{clip}(g) + N(0, \sigma^2 I) \quad (6)$$

This formula illustrates the noise-injection process. First, the gradient g is clipped to a maximum norm, and Gaussian noise N with variance σ^2 is added. This ensures that the model learns general patterns without overfitting to any data point.

Beyond privacy, Intellectual Property protection is critical because training state-of-the-art models costs millions. Model Extraction (Stealing) involves querying a victim model to train a surrogate "Knockoff Net," which can achieve high accuracy with a small amount of data. To combat this, owners use deep neural network (DNN) watermarking. In white box watermarking, the owner embeds a bitstring b into the weights W during training using a regularization term:

$$L_{total} = L_{task} + \lambda \cdot L_{watermark}(W, b) \quad (7)$$

This loss function combines the standard task loss L_{task} with a watermark-specific loss $L_{watermark}$, which is weighted by λ. The watermark loss forces the model's weights to conform to a specific statistical pattern (signature b) that can be mathematically verified later to prove ownership.

Fingerprinting is used for black box scenarios where weights are inaccessible. The model is trained to exhibit deliberate errors on a "Trigger Set" (e.g., classifying a specific noisy cat image as a "Toaster"). If a suspected stolen model replicates this specific error, it serves as strong legal evidence of the theft. Additionally, proof of training (PoT) protocols provides a "digital receipt" of the compute resources expended.

The era of Generative AI and Large Language Models (LLMs) introduces new risks. Prompt Injection acts as the "SQL Injection" of the NLP world, where attackers craft inputs to bypass system instructions. This can be Direct Injection (jailbreaking), such as "Persona Adoption" to bypass safety filters, or Indirect Prompt Injection, where the attack vector is hidden in data consumed by the LLM (e.g., invisible text in an email). Furthermore, hallucinations and supply chain poisoning pose severe threats; attackers can induce models to hallucinate non-existent software packages and then create malicious versions of those packages to compromise developers who rely on the generated code.

Moving to the edge of the network, the Internet of Things (IoT) presents a massive attack surface with billions of devices. Paracha et al. [1] highlights that standard security measures, such as heavy encryption and large anti-virus agents, are infeasible for battery-powered sensors because of resource constraints. Edge devices often operate on milliwatts of power with limited random-access memory (RAM), making it impossible to run a full deep learning-based intrusion detection system (IDS). To address this, the zero-watermarking-based data provenance in the IoT networks (ZIRCON) framework

proposes a "zero-watermarking" architecture. Unlike traditional watermarking, which modifies data (and risks corrupting sensor readings), this approach extracts intrinsic features from the data signal itself (e.g., a hash or feature vector V = f (D)) to generate a "logical watermark." This vector is XORed with a secret key and sent to a Provenance Database for verification, ensuring data provenance without a heavy overhead. This is coupled with a ZW-IDS (intrusion detection), which uses lightweight classifiers at the edge to quickly discard packets that lack a valid provenance trace, effectively filtering out spoofed data injection attacks.

Finally, the "black box" nature of deep learning remains a primary security vulnerability. If we do not know why a model made a decision, we cannot differentiate between a genuine error and an adversarial attack. Explainable AI (XAI) serves as a critical security primitive [9]. discusses two primary tools: LIME (local interpretable model-agnostic explanations), which approximates complex models with simple linear ones locally, and SHAP (Shapley additive explanations), which uses game theory to assign importance values to features. These tools are essential for detecting "Clever Hans" vulnerabilities, where models learn spurious correlations (e.g., a pentagon algorithm detecting tanks based on "cloudy days" rather than the tank itself). XAI allows security analysts to audit the reasoning of the model, not just the output, ensuring robustness before deployment.

FUTURE WORKS

Shoib et al. [5] provided a speculative roadmap for the evolution of AI security over the next 50 years.

In the 2025–2035: The integration era, we expect ubiquitous deployment of AI Copilots in coding, law, and medicine. The primary threat involves automated phishing and "deepfake" social engineering. Defense strategies will rely on standardized "Red Teaming" mandates and "Model Bills of Materials" (MBOM) to track supply chain ingredients.

Moving to the 2035–2050: The autonomy era, the focus shifts to artificial general intelligence (AGI) emergence, with systems capable of general reasoning. This threat has escalated to self-modifying malware powered by AI. Defense will require "self-healing" networks that patch their own vulnerabilities in real time, faster than human intervention is possible.

Finally, in the 2050–2075: The symbiosis era, we envision sentient systems in which AI manages planetary-scale resources (climate and energy grid). The security paradigm shifts from "defense" to "alignment." The primary challenge is ensuring that the AI's utility function remains aligned with human survival. Global governance frameworks, such as nuclear non-proliferation treaties, will be required for high-capability AI.

Securing AI is a task that must be solved not only by the technological world but also by the economic world. Hence, the threat posed by the "AI Trust Gap" will hinder the use of this technology. If users cannot trust the privacy of their data, as posed by model inversion attacks, or the output accuracy, as posed by hallucinations, then adoption into vital sectors will be halted.

The European Union's AI Act distinguishes its classification of AI systems according to their risk levels. Highly risky areas, such as "biometrical identification" or "critical infrastructure," must undergo strict assessments for conformity, including adversarial robustness. This rule compels firms to transform their development mantra from "Move Fast and Break Things" to "secure-by-design." This led to the development of a new market referred to as "AI assurance," which essentially involves third-party auditors checking a model's resistance to FGSM attacks and its adherence to the DP benchmark before being allowed to go live.

Finally, concerning the cost of security, applying tools based on DP and adversarial training incurs costs. Adversarial training is expected to result in a 3–10x increase in training time. This may also

impair the performance of clean data. Notably, the risk of attack must be measured against the cost of defense and the cost of utility.

CONCLUSION

The future of AI is undoubtedly revolutionary, but the threat landscape associated with it is more complex than ever. Starting from the pixel-level noise of the clean-label backdoor attack to the meaning manipulation in prompt injection, the 'AI attack surface' is massive and growing. Based on the above analysis, this study proposes the following research questions regarding the new characteristics and qualities that an AI system should possess. This study proves that the "security" feature of AI is a feature that cannot be incorporated after the fact. This feature must be inherent. Data integrity involves extensive tracking of provenance (zero-watermarking), model privacy requires the mathematical assurance of DP, and model trust requires explainability (XAI), which enables the auditing of the decision-making process. With the progression towards the age of autonomous and sentient systems, the role of vital cybersecurity professionals will transition from being network gatekeepers to algorithm integrity protectors. The future will belong to "secure-by-design" system architects who envision the attack characteristics inherent in the nature of intelligence.

REFERENCES

1. Paracha A, Arshad J, Farah MB, Ismail K. Machine learning security and privacy: a review of threats and countermeasures. EURASIP J Inf Secur. 2024;2024(1):10. doi:10.1186/s13635-024-00158-3.
2. Mohanty SS, Tripathy S. Application of different filtering techniques in digital image processing. J Phys Conf Ser. 2021;2062(1):012007. doi:10.1088/1742-6596/2062/1/012007.
3. Patil P, Chaudhary N, Prasad S, Bhandwal M, Arora M, Singh G. Predicting student performance with machine learning algorithms. 2023 3rd International Conference on Technological Advancements in Computational Sciences (ICTACS), Tashkent, Uzbekistan, 2023. p. 1346–1353. doi:10.1109/ICTACS59847.2023.10390077.
4. Xue B, Zhang M, Browne WN. Particle swarm optimization for feature selection in classification: a multi-objective approach. IEEE Trans Cybern. 2013;43(6):1656–1671. doi:10.1109/TSMCB.2012.2227469.
5. Shoib S, Siddiqui MF, Turan S, Chandradasa M, Armiya'u AY, Saeed F, et al. Artificial intelligence, machine learning approach and suicide prevention: a qualitative narrative review. Prev Med Res Rev. 2025;2(6):289–297. doi:10.4103/PMRR.PMRR_121_24.
6. Wang T, Zhang Y, Qi S, Zhao R, Xia Z, Weng J. Security and privacy on generative data in AIGC: a survey. ACM Comput Surv. 2024;57(4):1–34. doi:10.1145/3703626.
7. Zhang Y, Zeng D, Luo J, Xu Z, King I. A survey of trustworthy federated learning with perspectives on security, robustness and privacy. In: Ding Y, Tang J, Sequeda J, et al., editors. Companion Proceedings of the ACM Web Conference 2023 (WWW '23 Companion); 2023 Apr 30–May 4; Austin, TX, USA. New York (NY): Association for Computing Machinery; 2023. p. 1167–1176. doi:10.1145/3543873.3587681.
8. Rupanetti D, Kaabouch N. Combining edge computing-assisted internet of things security with artificial intelligence: applications, challenges, and opportunities. Appl Sci. 2024;14(16):7104. doi:10.3390/app14167104.
9. Wang H, Lv T, Cao Y, Li W, Zeng J, Huang P, et al. Navigating the dual-use nature and security implications of reconfigurable intelligent surfaces in next-generation wireless systems. IEEE Commun Surv Tutor. 2026;28:3346–3387. doi:10.1109/COMST.2025.3621610.
10. Prajapati T. Securing the AI ecosystem: a deep dive into mobile threats, countermeasures, and privacy-preserving techniques. Artif Intell Mob Comput. 2024;1(1).
11. 11. Liu X, Qian C, Hatcher WG, Xu H, Liao W, Yu W. Secure Internet of Things (IoT)-based smart-world critical infrastructures: survey, case study and research opportunities. IEEE Access. 2019;7:79523–79544. doi:10.1109/ACCESS.2019.2920763.
12. Aouedi O, Vu TH, Sacco A, Nguyen DC, Piamrat K, Marchetto G, et al. A survey on intelligent Internet of Things: applications, security, privacy, and future directions. IEEE Commun Surv Tutor. 2025;27(2):1238–1292. doi:10.1109/COMST.2024.3430368.

