Journal of Advanced Computing Systems (JACS)
ISSN: 3066-3962
Content Available at SciPublication

Adversarial Machine Learning in Digital Payments: A Framework for Detecting and Mitigating Evasion and Poisoning Attacks

Utham Kumar Anugula Sethupathy[1], *Vijayanand Ananthanarayanan*[2]
[1]*Independent Researcher, Senior IEEE Member, Alumni, Nanyang Technological University, Atlanta, USA*
[2]*Independent Researcher, Alumni, Fairleigh Dickinson University, Atlanta, USA*
DOI: 10.69987/JACS.2024.41207

Keywords

Artificial Intelligence, Adversarial Machine Learning, MITRE ATLAS, AI Security

Abstract

While Artificial Intelligence (AI) offers unprecedented capabilities for fraud detection and risk assessment in the digital payments ecosystem, the AI models themselves have emerged as a new, high-value attack surface. This paper provides a comprehensive analysis of the emerging threat landscape of adversarial machine learning (AML) in finance, with a specific focus on evasion, data poisoning, and model extraction attacks that can undermine the integrity of payment systems. We argue that traditional cybersecurity controls are insufficient to protect AI systems from these unique threats. To address this gap, we propose a comprehensive AI Resilience Framework for financial services. This framework integrates governance principles from the NIST AI Risk Management Framework (RMF) and MITRE ATLAS, specifies a secure ML-pipeline architecture (MLSecOps), details defense-in-depth mechanisms such as adversarial training, and outlines a robust program for adversarial testing and red teaming. This framework provides a practical, structured roadmap for financial institutions to build secure, robust, and trustworthy AI systems capable of withstanding sophisticated adversarial manipulation.

1. Introduction: The New Frontier of AI Security

The integration of Artificial Intelligence into financial services has fundamentally altered the security landscape. As institutions become increasingly reliant on AI for critical functions like real-time fraud detection, credit scoring, and algorithmic trading, adversaries are shifting their focus. Instead of solely targeting traditional infrastructure like networks and servers, they are now developing techniques to attack the AI models directly.[110] This has given rise to the field of adversarial machine learning (AML), which involves the intentional manipulation of AI systems to cause them to make incorrect predictions, reveal sensitive information, or otherwise behave in unintended ways.[112]

These attacks pose a profound threat to the financial sector. A successful adversarial attack could cause a fraud detection system to misclassify fraudulent transactions as legitimate, enable unqualified applicants to receive loans, or manipulate trading algorithms to cause market disruption. Securing against these threats requires a new security paradigm that extends beyond traditional application security. This new approach, often termed MLSecOps (Machine Learning Security Operations), must address the unique vulnerabilities present in the end-to-end AI/ML lifecycle, from data acquisition to model deployment and monitoring.[63] This paper introduces a resilience framework designed to provide financial institutions with a structured approach to this new security challenge.

2. The Threat Landscape of Adversarial ML in Finance

Adversarial attacks on machine learning systems can be categorized into several distinct types, each with specific implications for the financial industry. As illustrated in Figure 1, adversarial machine learning attacks in digital payment ecosystems can be broadly categorized into evasion attacks, data poisoning attacks, model extraction attacks, and privacy inference attacks. Each attack type targets a different stage of the AI lifecycle and introduces distinct operational and financial risks for institutions relying on AI-driven decision systems.

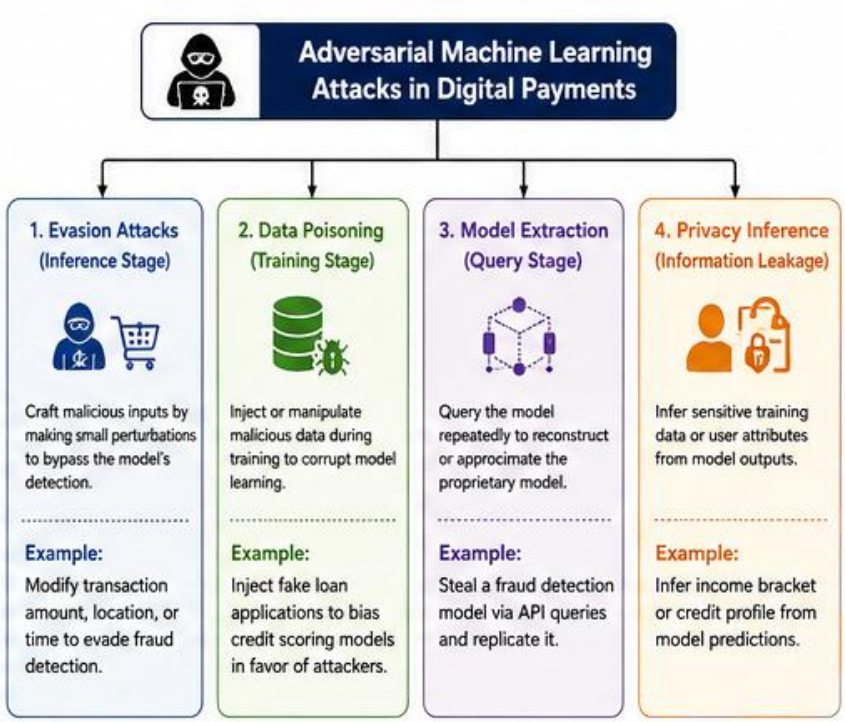

Figure 1. Taxonomy of Adversarial Machine Learning Attacks in Financial Payment Systems

2.1 Evasion Attacks: Bypassing Fraud Detection Models

Description: Evasion is the most common type of adversarial attack, occurring at the inference stage (when the model is making predictions). The attacker crafts a malicious input by making small, often imperceptible perturbations to a legitimate input, with the goal of causing the model to misclassify it.

Financial Example: A fraudster intends to use a stolen credit card for an online purchase. They know the bank uses an AI model to detect fraud. Before submitting the transaction, the attacker uses an evasion technique to slightly modify the transaction data—perhaps by minutely altering the purchase amount, using a VPN to spoof the location, or changing the time of day. These changes are small enough to appear legitimate but are specifically designed to push the transaction data across the model's decision boundary, causing the AI system to classify the fraudulent transaction as benign and approve it.

2.2 Data Poisoning Attacks: Corrupting Credit Scoring and Risk Models

Description: Data poisoning is a more insidious attack that targets the model's training phase. The attacker contaminates the training dataset with malicious examples, which corrupts the learning process and can create a "backdoor" in the model or degrade its overall performance.[113] Due to the reliance on vast, often publicly sourced datasets and the difficulty of detection, data poisoning is considered one of the most significant security threats to modern ML systems.[116]

Financial Example: An attacker wants to secure loans for a group of unqualified individuals. Over several months, they find a way to inject carefully crafted, falsified loan application data into a bank's data pipeline (e.g., through a compromised third-party data provider). This "poisoned" data trains the bank's credit scoring model to learn incorrect patterns, such as associating a specific postal code or a particular type of employment with high creditworthiness, regardless of other risk factors. When the attacker's associates later apply for loans using these characteristics, the compromised model incorrectly assigns them high credit scores, leading the bank to approve high-risk loans.

2.3 Model Extraction and Privacy Inference Attacks

Description: These attacks exploit the model as an information oracle. In a model extraction (or model stealing) attack, the adversary repeatedly queries a black-box model to reconstruct a functional copy of the

proprietary model itself. In a privacy inference attack, the goal is to infer sensitive information about the private data on which the model was trained.

Financial Example: A rival hedge fund could use model extraction techniques against a proprietary algorithmic trading model. By sending a large number of queries and observing the model's predictions, they can reverse-engineer the trading strategy, eroding the original firm's competitive advantage. In a privacy inference scenario, an attacker could probe a health insurance premium pricing model to determine if a specific individual (e.g., a public figure) was part of a high-risk group in the training data, thereby inferring sensitive health information.

3. A Resilience Framework for Financial AI Systems

Defending against adversarial attacks requires a holistic, defense-in-depth framework that integrates governance, secure architecture, and continuous validation. We propose a framework built on a foundational principle: a **Zero-Trust AI Supply Chain**. Modern AI systems are not built from scratch; they are assembled from a complex supply chain of components, including open-source libraries, pre-trained foundation models, and third-party data APIs.[111] Each of these external dependencies is a potential vector for attack. A compromised dataset can enable data poisoning; a backdoored pre-trained model can compromise every system that uses it.[63] Therefore, a secure financial AI posture cannot implicitly trust any component. It must adopt a "never trust, always verify" approach, where every artifact—data, model, and code—is authenticated, scanned, and monitored throughout its lifecycle, regardless of its origin.

Figure 2 illustrates the architectural design of the proposed resilient financial AI system, built on the principle of a zero-trust AI supply chain. The framework integrates governance controls, secure MLSecOps practices, runtime monitoring, and adversarial testing to establish a comprehensive defense posture for high-risk financial AI applications.

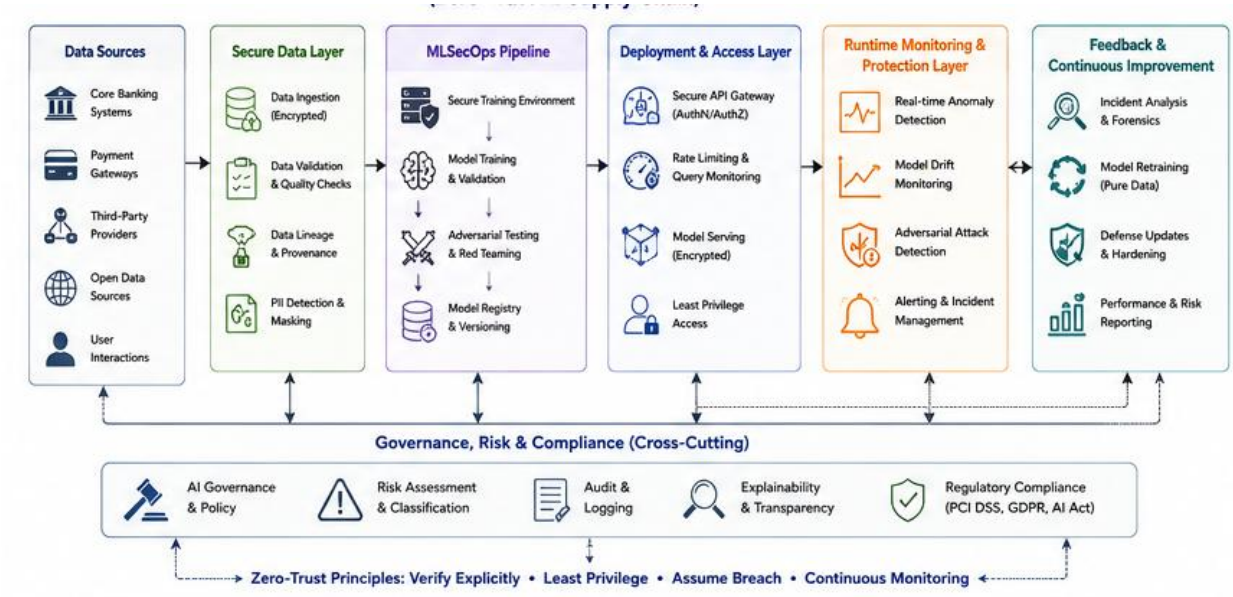

Figure 2. Architectural Diagram of a Resilient Financial AI System

3.1 Governance: Integrating NIST AI RMF and MITRE ATLAS

NIST AI Risk Management Framework (RMF): The NIST AI RMF provides the overarching governance structure. Financial institutions should use its four functions—**Govern, Map, Measure, and Manage**—to establish policies for AI risk, identify and map potential adversarial threats to specific business processes, define metrics to measure model robustness, and implement and manage controls to mitigate identified risks.

MITRE ATLAS Framework: While NIST provides the "what," MITRE ATLAS (Adversarial Threat Landscape for AI Systems) provides the "how." ATLAS is a knowledge base of adversary tactics and techniques based on real-world observations. Teams should use ATLAS to conduct detailed threat models, mapping specific adversary behaviors (e.g., T0029 - Poison Training Data, T0040 - Evade ML Model) to their financial AI systems. This creates a common vocabulary for discussing and defending against AI-specific attacks.

3.2 Secure AI/ML Pipeline Architecture (MLSecOps)

The AI Resilience Framework must be implemented through a hardened MLSecOps pipeline, as illustrated in Figure 3. Figure 3 presents the secure MLSecOps lifecycle adopted in the proposed AI Resilience Framework. The lifecycle emphasizes continuous validation across data ingestion, secure model training, adversarial testing, deployment hardening, runtime monitoring, and incident response, ensuring resilience throughout the operational lifespan of financial AI systems.

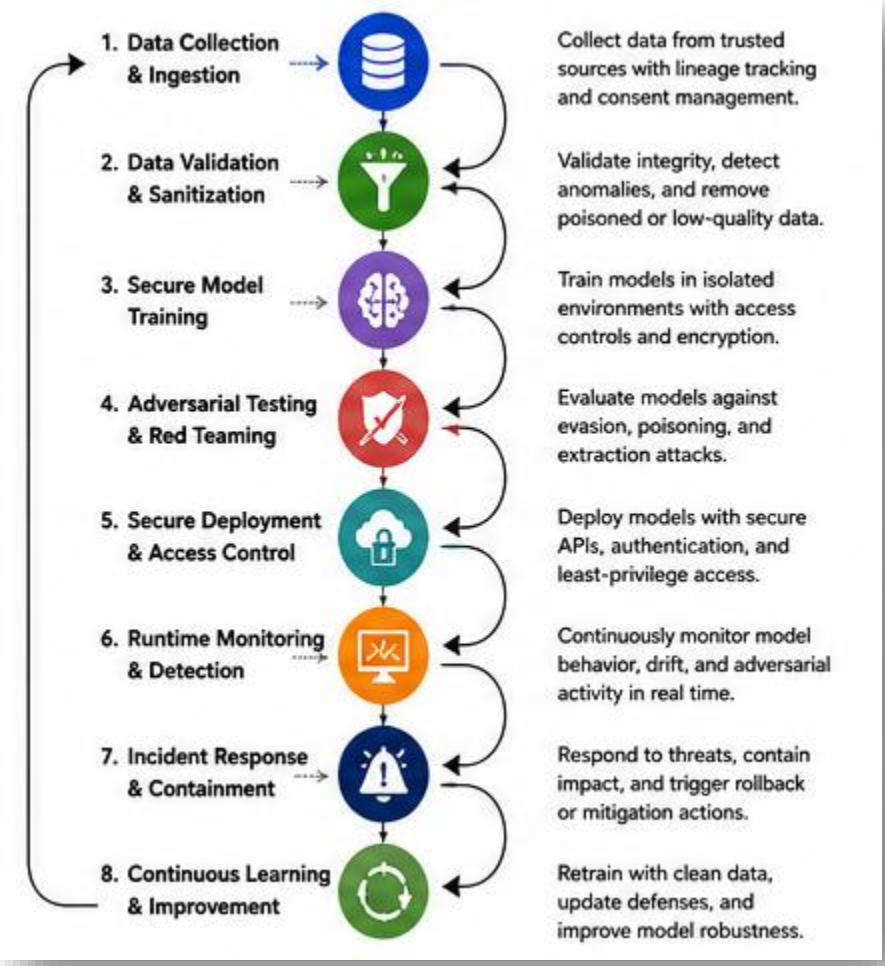

Figure 3. Secure MLSecOps Lifecycle for Financial AI Systems

Secure Data Ingestion and Preprocessing: The pipeline must begin with strong data validation, sanitation, and outlier detection mechanisms to defend against data poisoning attacks. All sensitive financial data must be encrypted in transit and at rest, and techniques like data masking or tokenization should be used to protect personally identifiable information (PII) in training data.

Robust and Secure Model Training: Model training should occur in secure, isolated environments. A key defense mechanism, **adversarial training**, should be implemented. This involves intentionally generating adversarial examples and including them in the training dataset, which teaches the model to be more robust against such perturbations.

Secure Model Deployment: Models should be treated as critical software artifacts. They must be digitally signed to ensure their integrity and encrypted to protect proprietary architectures. Deployments should be containerized (e.g., using Docker) with hardened configurations and minimal privileges. A versioned model registry (e.g., MLflow) is essential for traceability and enabling rapid rollbacks if an issue is detected.

3.3 Runtime Monitoring and Defense

Input and Output Validation: All inputs to a production model must be validated and sanitized to defend against inference-time attacks like prompt injection in LLMs or other evasion techniques. Similarly, model outputs should be filtered to prevent the leakage of sensitive or harmful information.

Anomaly and Data Drift Detection: Security is not a one-time check. Systems must continuously monitor the performance of deployed models and the statistical properties of the data they are processing. A sudden degradation in performance or a significant change in the input data distribution (known as "data drift") can be an indicator of an ongoing attack or a change in the environment that renders the model obsolete. Such events should trigger automated alerts for investigation and potential model retraining.[118]

Table 1 compares existing AI security frameworks commonly referenced in financial services, including NIST AI RMF, MITRE ATLAS, and traditional DevSecOps approaches. The comparison highlights that the proposed AI Resilience Framework provides broader coverage by integrating governance, adversarial defense, runtime monitoring, financial-domain specificity, and continuous red teaming into a unified operational model.

Table 1. Comparative Analysis of Existing AI Security Frameworks for Financial Services

Framework	Governance	Adversarial Defense	Runtime Monitoring	Financial Context	Red Teaming
NIST AI RMF	Yes	Partial	No	Generic	No
MITRE ATLAS	Partial	Yes	No	Generic	Yes
Traditional DevSecOps	Partial	No	Partial	Limited	No
Proposed AI Resilience Framework	Yes	Yes	Yes	Strong	Yes

4. Adversarial Testing and Red Teaming for Financial AI

To ensure that defenses are effective, they must be tested. A proactive adversarial testing program is a critical component of the resilience framework.

4.1 Tools and Techniques for Simulating Attacks

Financial institutions should establish an AI red team responsible for simulating adversarial attacks against their models. This team can leverage a suite of open-source and commercial tools to orchestrate these tests:

Adversarial Attack Frameworks: The **Adversarial Robustness Toolbox (ART)** from IBM and **Counterfit** from Microsoft are powerful, model-agnostic frameworks for generating a wide range of evasion and poisoning attacks.[121] Other libraries like **Foolbox** and **TextAttack** provide specialized capabilities for image and text-based models, respectively.

Specialized Testing Tools: For systems using LLMs, tools like **PentestGPT** and **Garak** can help automate the process of finding vulnerabilities like prompt injection.[121]

As shown in Figure 4, different categories of adversarial attacks require distinct defensive controls and monitoring strategies. A defense-in-depth approach that maps specific attack vectors to appropriate mitigation techniques significantly improves model robustness and reduces operational exposure.

4.2 Key Performance Indicators (KPIs) for Model Robustness

The success of a secure AI program cannot be measured by predictive accuracy alone. New security-specific KPIs are required to quantify model resilience:

Technical Performance Metrics: Standard metrics like Model Uptime, Prediction Latency, and Error Rate remain important for operational health.

Security-Specific Metrics:

Adversarial Success Rate: The percentage of adversarial examples generated by a red team that successfully cause the model to make an incorrect prediction. The goal is to minimize this rate.

Minimum Perturbation Distance: A measure of model robustness. This KPI quantifies the smallest amount of "noise" or change needed to be added to an input to cause a misclassification. A higher value indicates a more robust model.

Mean Time to Detect (MTTD) and Mean Time to Respond (MTTR) for AI Incidents: These metrics measure the efficiency of the monitoring and incident response processes for AI-specific threats.

Attack Type	Primary Goal	Key Defense Controls	Detection Mechanisms
Evasion Attacks	Bypass model detection at inference stage.	• Adversarial Training • Input Validation & Sanitization • Ensemble Models • Anomaly Detection	• Input Anomaly Detection • Behavior Analysis • Confidence Monitoring • Adversarial Example Detection
Data Poisoning	Corrupt training data to alter model behavior.	• Data Validation • Outlier Detection • Provenance & Lineage • Robust Training	• Data Drift Detection • Statistical Testing • Integrity Monitoring • Backdoor Detection
Model Extraction	Reconstruct or steal proprietary model.	• Rate Limiting • Query Auditing • Output Perturbation • Model Watermarking	• Query Pattern Analysis • Usage Anomaly Detection • Extraction Detection Tests
Privacy Inference	Infer sensitive information from outputs.	• Differential Privacy • k-Anonymity • Output Suppression • Access Controls	• Membership Inference Detection • Attribute Inference Detection • Privacy Budget Tracking
Prompt Injection*	Manipulate model behavior via malicious prompts (LLM systems).	• Input Sanitization • Prompt Filtering • Guardrail Policies • Context Isolation	• Prompt Anomaly Detection • Guardrail Monitoring • Response Validation

*Applicable for LLM-based payment assistants and conversational AI systems.

Figure 4. Mapping of Adversarial Attacks to Defensive Controls and Detection Mechanisms

Model Explainability Score: Metrics from XAI tools can be used to ensure that models remain transparent and auditable, which is itself a form of security control against unexplainable behavior.

5. Discussion: The Future of Secure and Trustworthy Financial AI

The field of adversarial machine learning is a dynamic "arms race," with researchers continuously discovering new attack vectors and developing corresponding defenses.[114] For the financial industry, achieving absolute, impenetrable security is an unrealistic goal. The strategic objective must instead be

resilience—the ability to withstand attacks, detect them quickly, and recover gracefully with minimal impact.

As the regulatory landscape evolves, particularly with frameworks like the EU AI Act, it is highly probable that requirements for formal adversarial robustness testing and transparent risk management will become standard practice for any high-risk financial AI system. Proactively adopting a comprehensive resilience framework now will not only mitigate current risks but also position institutions to be compliant with the regulations of the future.

6. Conclusion

The deployment of AI in digital payments and financial services has opened a new and critical front in cybersecurity. Adversarial attacks targeting the AI models themselves represent a sophisticated and potent threat that cannot be addressed by traditional security measures alone. The AI Resilience Framework proposed in this paper—built on the principles of a zero-trust supply chain, integrated governance, secure MLSecOps architecture, and continuous adversarial testing—provides a necessary evolution of security practice. By treating AI models not just as assets to be used but as systems to be defended, financial institutions

can move toward a future where AI is deployed not only effectively but also safely, securely, and responsibly, thereby maintaining the trust that is fundamental to the financial system.

6. References

[1] National Institute of Standards and Technology, *Artificial Intelligence Risk Management Framework (AI RMF 1.0)*, NIST AI 100-1, 2023.

[2] MITRE Corporation, *MITRE ATLAS™: Adversarial Threat Landscape for Artificial-Intelligence Systems*, 2024.

[3] OWASP Foundation, *OWASP DevSecOps Maturity Model (DSOMM)*, Version 2.0, 2024.

[4] IBM, N. Papernot et al., "The Limitations of Deep Learning in Adversarial Settings," in *Proc. IEEE European Symposium on Security and Privacy*, 2016.

[5] Google Research, I. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and Harnessing Adversarial Examples," *International Conference on Learning Representations (ICLR)*, 2015.

[6] N. Carlini and D. Wagner, "Towards Evaluating the Robustness of Neural Networks," in *Proc. IEEE Symposium on Security and Privacy*, 2017.

[7] B. Biggio and F. Roli, "Wild Patterns: Ten Years After the Rise of Adversarial Machine Learning," *Pattern Recognition*, vol. 84, pp. 317–331, 2018.

[8] M. Barreno, B. Nelson, A. Joseph, and J. Tygar, "The Security of Machine Learning," *Machine Learning*, vol. 81, no. 2, pp. 121–148, 2010.

[9] L. Huang, A. Joseph, B. Nelson, B. Rubinstein, and J. Tygar, "Adversarial Machine Learning," in *Proc. ACM Workshop on Security and Artificial Intelligence*, 2011.

[10] B. Nelson et al., "Near-Optimal Evasion of Convex-Inducing Classifiers," in *Proc. AISTATS*, 2010.

[11] S. Mei and X. Zhu, "Using Machine Teaching to Identify Optimal Training-Set Attacks on Machine Learners," in *Proc. AAAI Conference on Artificial Intelligence*, 2015.

[12] M. Jagielski et al., "Manipulating Machine Learning: Poisoning Attacks and Countermeasures for Regression Learning," in *Proc. IEEE Symposium on Security and Privacy*, 2018.

[13] F. Tramèr et al., "Stealing Machine Learning Models via Prediction APIs," in *USENIX Security Symposium*, 2016.

[14] R. Shokri et al., "Membership Inference Attacks Against Machine Learning Models," in *Proc. IEEE Symposium on Security and Privacy*, 2017.

[15] N. Papernot et al., "Practical Black-Box Attacks Against Machine Learning," in *Proc. ACM Asia Conference on Computer and Communications Security*, 2017.

[16] A. Madry et al., "Towards Deep Learning Models Resistant to Adversarial Attacks," in *International Conference on Learning Representations (ICLR)*, 2018.

[17] Microsoft, *Counterfit: A General Automation Layer for AI Security Testing*, 2024.

[18] IBM, *Adversarial Robustness Toolbox (ART)*, 2024.

[19] MLflow, *MLflow Model Registry and Secure Deployment Framework*, 2024.

[20] PCI Security Standards Council, *Payment Card Industry Data Security Standard (PCI DSS) Version 4.0*, 2022.

[21] European Union, *EU Artificial Intelligence Act*, 2024.

[22] Stripe, "How Machine Learning Works for Payment Fraud Detection and Prevention," Technical Whitepaper, 2024.

[23] IBM, "AI Fraud Detection in Banking," IBM Financial Services Whitepaper, 2024.

[24] A. Doshi-Velez and B. Kim, "Towards a Rigorous Science of Interpretable Machine Learning," arXiv preprint arXiv:1702.08608, 2017.

[25] S. M. Lundberg and S. I. Lee, "A Unified Approach to Interpreting Model Predictions," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

[26] M. T. Ribeiro, S. Singh, and C. Guestrin, "Why Should I Trust You? Explaining the Predictions of Any Classifier," in *Proc. ACM SIGKDD*, 2016.

[27] Microsoft Learn, *Threat Modeling AI/ML Systems and Dependencies*, 2024.

[28] National Cybersecurity Center of Excellence, *Artificial Intelligence: Adversarial Machine Learning*, NCCoE Practice Guide, 2024.

[29] Protect AI, *MLSecOps: The Foundation of AI/ML Security*, Technical Report, 2024.

[30] S. Singh et al., "DocLLM: A Layout-Aware Generative Language Model for Multimodal Document Understanding," in *Proc. ACL 2024*, pp. 8451–8468, 2024.