AI-Based Detection of Hedge Fund Fraud

Section 5 – Adversarial Robustness, Regulatory Readiness, and Ethics (C2)

Joerg Osterrieder

Zurich University of Applied Sciences (ZHAW)

2025

Outline

The Adversary Profile

Not Script Kiddies – PhD-Level Quants

- Adversaries are **quantitatively trained professionals** (PhDs in math, physics, CS)
- Deep knowledge of statistical methods and the same academic literature informing detection systems
- Access to sophisticated modeling tools, computational resources, and expert consultants
- Strong financial incentives to remain undetected ($\Rightarrow$ billions at stake)

Key Implication

Standard adversarial ML threat models (pixel perturbations, ℓ_p-norm budgets) are **insufficient**. The hedge fund adversary can:

- Simulate detection models
- Engineer statistically plausible returns
- Exploit temporal and structural gaps

78% of financial institutions lack formal adversarial resilience policies for ML-based detection.

Source: Goodfellow et al. (2015); Biggio & Roli (2018); paper Section 5.1

Attack Vector 1: Data Poisoning

Mechanism

Fraudulent managers report **fabricated returns** to commercial databases (HFR, Lipper TASS), effectively poisoning the training data of supervised detection models.

Why It Works

- Hedge fund reporting is **voluntary and largely unverified**
- A manager can engineer return series that:
  - Satisfy Benford's law
  - Exhibit appropriate serial correlation
  - Maintain plausible distributional properties
- Small fraction of poisoned samples $\Rightarrow$ disproportionate effect

Impact

- **5–12% model performance degradation** even with small poisoning fraction (Goldblum et al., 2023)
- Particularly dangerous because:
  - Base rate of fraud is already low
  - Class imbalance **amplifies** the impact of corrupted labels
- Undermines the foundational assumption of clean training data

Source: Goldblum et al. (2023); paper Section 5.1

Attack Vector 2: Evasion Attacks

Mechanism

Structuring reported returns to **avoid triggering detection thresholds** – the most natural adversarial attack in the hedge fund domain.

Forms of Evasion

- **Return smoothing** (Getmansky et al., 2004): suppresses volatility signals and serial correlation
- Optimizing reported returns against a known or estimated **detection boundary**
- Minor perturbations bounded by financial plausibility $\Rightarrow$ substantially alter predictions

Quantified Impact

- FGSM and PGD attacks degrade AUC by **8–15%** (Cartella et al., 2021)
- Mean AUC degradation across surveyed systems: **10.6%**
- Even minor plausibility-bounded perturbations:
  - Elevate calibration error
  - Increase expected portfolio loss by $\sim$5%

Source: Cartella et al. (2021); Goodfellow et al. (2015); Madry et al. (2018); paper Section 5.1

Attack Vector 3: Model Extraction

Mechanism

Reverse-engineering a regulator's detection model by **observing enforcement patterns** over time.

- A sophisticated adversary can infer which features and thresholds trigger scrutiny by analyzing:
  - Which funds are investigated
  - Which enforcement actions are brought
  - Which anomalies are flagged in examination letters
- Construct a **surrogate model** of the detection system
- Optimize reported behavior to remain below the decision boundary

Facilitating Factor

SEC enforcement actions and examination priority announcements are **public** – inadvertently revealing features and thresholds that regulators prioritize.

Source: Paper Section 5.1

Attack Vector 4: Strategic Timing and Regime Exploitation

Market Stress Exploitation

- Time fraudulent activity to coincide with **periods of market stress**
- During crises, legitimate fund returns exhibit unusual distributional properties
- Fabricated returns are **masked within broader market noise**
- Detection models struggle to distinguish fraud from genuine market dislocation

Gradual Introduction

- Introduce fraudulent reporting **incrementally**
- Detection models trained on historical data treat evolving fraud as **legitimate regime change**
- Exploits the assumption of stationarity in most supervised models
- By the time the pattern becomes detectable, years of damage may have occurred

$\Rightarrow$ Most detection models do not account for these temporal dynamics – a critical vulnerability.

Source: Paper Section 5.1

Attack Vectors: Summary

- All four vectors exploit structural features of the hedge fund ecosystem
- Evasion and data poisoning have the most **empirically quantified** impact
- Model extraction and strategic timing are harder to measure but equally dangerous

Source: Paper Section 5.1

Mean AUC Degradation: 10.6%

Aggregate Finding

Across surveyed detection systems, adversarial perturbations cause a **mean** AUC **degradation of 10.6%**, with some techniques exhibiting degradation **exceeding 25%** under informed adversaries.

Breakdown

- Data poisoning: **5–12%** degradation
- Evasion (FGSM/PGD): **8–15%** degradation
- Plausibility-bounded perturbations: ∼**5%** portfolio loss increase
- Attack success rate (pre-defense): ∼**35%**

Why This Understates the Risk

- Literature tests *generic* perturbations, not financially tailored attacks
- Real adversaries use **domain-specific constraints**
- Actual degradation against PhD-level quants likely **higher**
- No published study tests robustness against hedge-fund-specific adversaries

⇒ Detection performance on historical data is necessary but **not sufficient** for deployment.

Source: Cartella et al. (2021); Chen et al. (2024); paper Section 5.1

Defense 1: Adversarial Training

Approach

Augment training data with adversarial examples, forcing the model to learn decision boundaries **robust to perturbations** (min-max optimization: Madry et al., 2018).

Effectiveness

- Recovers **60–70%** of AUC lost to adversarial attacks (Chen et al., 2024)
- Reduces attack success rate from ∼35% to ∼**5%**
- Most direct and well-studied defense mechanism

Limitations

- Computationally expensive: **5–10**× standard training cost
- Which perturbation model to use?
  - ℓ_∞-bounded perturbations dominate CV literature
  - May not capture **financially constrained** perturbations
- Need domain-specific perturbation budgets

Source: Madry et al. (2018); Chen et al. (2024); paper Section 5.2

Defense 2: Ensemble Diversity

Implicit Adversarial Robustness

Ensembles aggregating predictions from **multiple heterogeneous base learners** provide a natural form of adversarial robustness.

- An adversary optimizing evasion against one model component is **unlikely to simultaneously fool all components**
- Most effective when ensemble incorporates **diverse model families**:
  - Tree-based models (XGBoost, LightGBM)
  - Neural networks (LSTM, autoencoders)
  - Statistical anomaly detectors (isolation forests, one-class SVM)
- Diversity metric over ensemble components **correlates positively** with robustness (Patel et al., 2025)
- Yields robustness benefits **beyond** the accuracy gains typically reported for ensembles

$\Rightarrow$ Combining gradient-boosted trees with neural network classifiers is both a performance and a robustness strategy.

Source: Patel et al. (2025); paper Section 5.2

Defense 3: Input Validation and Meta-Level Anomaly Detection

Layered Defense

Deploy a **separate anomaly detection system** to screen model inputs for signs of adversarial manipulation – treating adversarial detection as a distinct task from fraud detection.

What the Meta-Detector Flags

- Suspiciously precise conformity to Benford's law
- Implausibly low serial correlation for the stated strategy type
- Distributional characteristics inconsistent with reported asset class exposures
- Statistical properties "too clean" to be genuine

Advantages

- Does not require modifying the classification model itself
- Enables **specialized models** for each task
- Can be updated independently as new attack patterns emerge
- Complements rather than replaces adversarial training

Source: Paper Section 5.2

Defense 4: Certified and Randomized Defenses

Approach

Certified defense methods (e.g., **randomized smoothing**) provide **provable robustness guarantees** within a specified perturbation radius.

Theoretical Appeal

- Formal guarantee: bounded input change ⇒ bounded output change
- Eliminates need to enumerate all possible attacks
- Provides **worst-case** robustness assurance

Practical Challenges

- Perturbation model must be defined over **financially meaningful dimensions** (returns, risk metrics, factor exposures)
- Certification radius must align with plausible adversarial strategies, not arbitrary ℓ_p norms
- No published work has adapted certified defenses to hedge fund fraud detection

⇒ An important **open area** for future research (see OP8).

Source: Paper Section 5.2

EU AI Act: Requirements for Fraud Detection Systems

Classification

Under Annex III of the EU AI Act (Regulation 2024/1689), AI systems for financial fraud detection are classified as **high-risk AI systems**, subject to the Act's most stringent requirements.

Art. 13: Transparency

- Operations must be **sufficiently transparent**
- Users must interpret output and use it appropriately
- $\Rightarrow$ Favors interpretable models or robust post-hoc explanations

Art. 14: Human Oversight

- Effective oversight by **natural persons**
- Ability to override output or intervene
- $\Rightarrow$ Codifies **human-in-the-loop** principle

Art. 9: Risk Management

- Comprehensive risk management system
- Must address **adversarial vulnerabilities**
- $\Rightarrow$ Links adversarial robustness to **legal compliance**

Entered into force August 2024, phased implementation through 2027.

Source: EU AI Act (2024/1689); paper Section 5.3.1

SEC Regulatory Expectations

Current Landscape

- No comprehensive AI legislation comparable to EU AI Act
- However, SEC signals **increasing regulatory attention**:
  - Division of Examinations: AI and emerging technology as **priority area**
  - Staff guidance: fiduciary obligations for algorithmic tools
- DERA itself employs ML models for market surveillance and enforcement targeting
- Creates **implicit benchmarks** for industry practice

AI Washing Enforcement

- Recent enforcement actions against firms that **exaggerated or fabricated AI use** in investment processes
- Signals SEC views **AI governance** as within enforcement purview
- Industry should anticipate:
  - More specific expectations for AI-based compliance systems
  - Governance requirements for ML-based surveillance

$\Rightarrow$ Even without legislation, SEC expectations will shape deployment standards.

Source: SEC Examination Priorities (2023); paper Section 5.3.2

The Explainability–Performance Trade-off

The Fundamental Tension

- Most accurate methods (deep ensembles, transformers, GNNs) are the **most opaque**
- Inherently interpretable models (logistic regression, decision trees, rules) offer transparency but **lower detection rates**
- Regulatory requirements favor transparency
- Detection requirements favor complexity

Post-Hoc Explainability Methods

- **SHAP** (Lundberg & Lee, 2017): theoretically grounded feature attribution scores
- **LIME** (Ribeiro et al., 2016): local surrogate models
- Can generate: "flagged due to low serial correlation, high Sharpe vs. peers, irregular distributions"
- Limitations: added cost, may not faithfully represent true decision process, can be unstable

Open Question

No regulatory consensus on what constitutes "sufficient" explainability for financial fraud detection AI. Compliance requirements may tighten retroactively.

Source: Rudin (2019); Lundberg & Lee (2017); Ribeiro et al. (2016); paper Section 5.3.3

Explainability vs. Performance Spectrum

ethod families (linear/logistic, tree ensembles, neural networks, hybrid architectures) along axes of detection performance (y) vs. explai

- Tree-based ensembles currently occupy the **most favorable position**: strong performance with moderate explainability via SHAP
- Hybrid architectures (tree classification + neural feature extraction) offer promising trade-offs

Source: Paper Sections 5.3.3 and 5.4

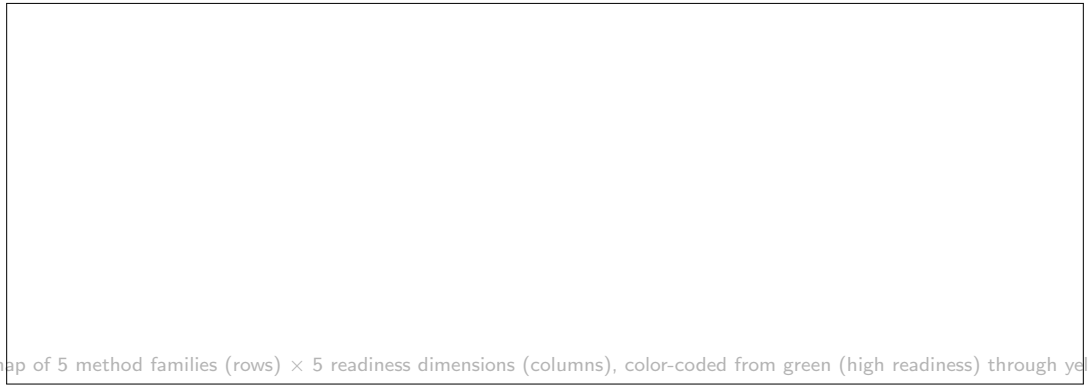
Readiness Assessment: 5 Methods × 5 Dimensions

Method Family	Adversarial Robustness	Intrinsic Explain.	Post-Hoc Explain.	Regulatory Readiness	Deployment Maturity
Linear / Logistic	Low	High	High	High	High
Tree-Based Ensembles	Moderate	Moderate	High	High	High
Deep Learning	Low	Low	Moderate	Low	Moderate
Hybrid / Ensemble	Mod–High	Moderate	Mod–High	Moderate	Moderate
Anomaly Detection	Moderate	Variable	Variable	Moderate	Moderate

- **Tree-based ensembles** (XGBoost, LightGBM, RF) occupy the most favorable overall position
- **Deep learning** offers highest potential accuracy but ranks lowest on explainability and regulatory readiness
- **Hybrid architectures** provide pragmatic path: interpretable classification + neural feature extraction

Source: Paper Section 5.4

Readiness Heatmap: Visual



- No single method family dominates across **all** criteria
- Selection depends on organizational priorities: accuracy vs. compliance vs. robustness

Source: Paper Section 5.4

Ethics: False Positives and Selection Bias

Harm from False Positives

- Consequences are **substantial and asymmetric**:
  - Severe reputational damage
  - Investor redemptions
  - Counterparty relationship termination
  - Regulatory scrutiny cascade
- Can **destroy a legitimate business** even if allegation is unfounded
- Market-level effects possible (contagion from public investigation)
- Categorically different from false positives in spam filtering

Selection Bias from Enforcement Data

- Supervised models trained on past sec *actionslearnpatternsof* **historically prosecuted** *fraud*
- Prosecuted cases $\neq$ random sample of all fraud
- Historical concentration on certain:
  - Fund types and strategies
  - Geographies
  - Visible fraud schemes
- Models may **perpetuate and amplify** selective scrutiny
- Under-detect in overlooked categories

Source: Paper Section 5.5

Ethics: Fairness, Dual-Use, and the Transparency Paradox

Fairness Across Fund Characteristics

- Do models disproportionately flag:
  - Small funds?
  - Emerging market funds?
  - Certain demographic groups?
- Almost **no empirical attention** in literature
- Need fairness metrics for financial context

Dual-Use Nature

- Same AI techniques used by regulators can be **repurposed by fraudsters**
- Published models provide a **roadmap for evasion**
- Not hypothetical: hedge fund quants consume and operationalize published research

Transparency Paradox

- Regulatory demands for disclosure **conflict with** operational security
- Every piece of disclosed information is exploitable by adversaries
- Current regulations do not adequately address this tension

$\Rightarrow$ Balancing transparency with adversarial security requires nuanced governance frameworks.

Source: Doshi-Velez & Kim (2017); paper Section 5.5

Governance Recommendations

Recommended Governance Framework

Organizations deploying AI-based fraud detection should adopt:

1. **Regular bias audits** across fund characteristics and demographic dimensions
2. **Diverse and representative training data** that corrects for historical enforcement biases where possible
3. **Human-in-the-loop decision-making**: AI systems inform but do **not replace** human judgment in enforcement decisions
4. **Transparent model governance**: documented procedures for model development, validation, monitoring, and retirement

- Aligns with emerging best practices in responsible AI (Arrieta et al., 2020; Molnar, 2020)
- Aligns with EU AI Act human oversight requirements (Art. 14)
- Adversarial testing should be **standard practice**, not optional

Source: Arrieta et al. (2020); Molnar (2020); paper Section 5.5

Summary: Adversarial Robustness, Regulatory Readiness, and Ethics

1. **Adversary profile**: PhD-level quants with deep statistical knowledge – not generic attackers
2. **Four attack vectors**: data poisoning (5–12%), evasion (8–15% AUC), model extraction, strategic timing
3. **Mean AUC degradation**: 10.6% across surveyed systems – likely understates real-world risk
4. **Defenses**: adversarial training (recovers 60–70%), ensemble diversity, input validation, certified methods (open area)
5. **EU AI Act**: fraud detection classified as high-risk; mandates transparency, human oversight, risk management
6. **SEC**: no comprehensive legislation yet, but increasing attention and implicit expectations
7. **Explainability–performance trade-off**: tree-based ensembles best positioned; no regulatory consensus on "sufficient" explainability
8. **Readiness**: no single method dominates all 5 dimensions; hybrid architectures most pragmatic
9. **Ethics**: false positive harm is severe, selection bias is systematic, dual-use and transparency paradox require governance
10. **Governance**: bias audits, diverse training data, human-in-the-loop, transparent model governance

Source: Paper Section 5 (Contribution C2)