AI-Based Detection of Hedge Fund Fraud: A Systematic Survey and Research Agenda

Joerg Osterrieder

Zurich University of Applied Sciences

School of Engineering

Switzerland

joerg.osterrieder@zhaw.ch

February 14, 2026

Abstract

The hedge fund industry, managing over $4.5 trillion in assets under management, faces persistent fraud risks amplified by operational opacity and limited regulatory oversight. Despite growing regulatory scrutiny and technological advances, no systematic survey has examined artificial intelligence methods specifically for hedge fund fraud detection. This paper addresses this gap through three contributions. First, we develop a unified five-stage detection pipeline taxonomy (data ingestion, feature engineering, model selection, explainability, deployment) that maps hedge fund fraud types to appropriate AI methods—the first hedge-fund-specific detection framework in the literature. Second, we conduct a systematic adversarial and regulatory readiness assessment, revealing significant vulnerabilities (mean AUC degradation of 10.6% under adversarial attack) and uncertain compliance with the EU AI Act and SEC requirements. We evaluate defense mechanisms and their effectiveness across different attack scenarios. Third, we propose a research agenda articulating ten concrete open problems spanning data challenges (benchmark datasets, cross-jurisdictional integration, real-time pipelines), methodological challenges (extreme

class imbalance with only 50–100 documented fraud cases, cold-start detection, concept drift, multi-modal fusion), and deployment challenges (adversarial robustness, explainability, human-AI collaboration). Our analysis shows that ensemble methods combining tree-based models with anomaly detection currently offer the most robust performance, yet critical gaps remain in adversarial resilience and standardized evaluation datasets. These findings have direct implications for practitioners designing fraud detection systems, regulators evaluating AI-based compliance tools, and researchers advancing financial crime prevention methodologies.

Keywords: hedge fund fraud, machine learning, anomaly detection, financial regulation, explainable AI, systematic review, adversarial robustness

JEL Classification: G23 (Non-bank Financial Institutions), G28 (Government Policy and Regulation), C45 (Neural Networks and Related Topics), C53 (Forecasting and Prediction Methods), K22 (Business and Securities Law)

Contents

1 Introduction

1.1 The Scale of Hedge Fund Fraud

The global hedge fund industry manages assets exceeding $4.5 trillion as of 2025, having grown substantially from roughly $2 trillion at the onset of the 2008 financial crisis. This expansion has been accompanied by an equally striking growth in the complexity and diversity of investment strategies, ranging from quantitative statistical arbitrage to activist equity positions and illiquid credit. Yet unlike mutual funds, which operate under the transparency and reporting requirements of the Investment Company Act of 1940, hedge funds have historically benefited from broad exemptions that permit limited disclosure, voluntary performance reporting, and minimal portfolio-level transparency (Stulz, 2007). These structural features—while enabling the strategic flexibility that attracts institutional capital—simultaneously create an environment in which fraud can persist undetected for years or even decades.

The financial toll of hedge fund fraud is enormous. The collapse of Bernard L. Madoff Investment Securities in December 2008 revealed a Ponzi scheme with estimated losses of $65 billion in stated account value, making it the largest financial fraud in history (**?**). The Bayou Group, a Connecticut-based hedge fund, concealed roughly $450 million in trading losses through fabricated financial statements and a sham auditing firm between 2003 and 2005. More recently, the implosion of Archegos Capital Management in March 2021 generated over $10 billion in counterparty losses across major prime brokers, exposing failures in risk monitoring and concentrated position reporting that regulators had not anticipated. These are not isolated incidents. The U.S. Securities and Exchange Commission (SEC) brings dozens of enforcement actions against hedge fund managers and private fund advisers each year, with violations spanning return misrepresentation, asset misappropriation, insider trading, and valuation manipulation.

Several characteristics render hedge funds uniquely vulnerable to fraudulent activity. First, hedge funds frequently invest in illiquid or hard-to-value assets—including distressed debt, private equity co-investments, and bespoke derivatives—for which indepen-

dent pricing is difficult or impossible to obtain (Getmansky et al., 2004). This opacity in valuation creates opportunities for managers to inflate reported net asset values (NAVs), smooth returns to conceal volatility, or fabricate performance altogether. Second, hedge fund reporting to commercial databases such as Hedge Fund Research (HFR), Lipper TASS, and Morningstar is entirely voluntary, introducing well-documented survivorship, backfill, and self-selection biases that complicate statistical analysis (Agarwal et al., 2011; Fung and Hsieh, 2009). Third, lock-up periods and redemption gates restrict investor liquidity, delaying the discovery of fraud by preventing investors from withdrawing capital when suspicions arise. Fourth, the limited partnership structures typical of hedge funds concentrate decision-making authority in a small group of general partners, often with minimal independent oversight. Taken together, these features create what Stulz (2007) characterize as an agency problem of unusual severity: managers possess both the incentive and the means to misrepresent fund performance, while investors and regulators lack the information and access needed for effective monitoring.

1.2 Why AI? The Limitations of Traditional Detection

The mismatch between regulatory capacity and industry scale is stark. The SEC employs approximately 4,600 staff to oversee thousands of registered investment advisers, broker-dealers, and fund complexes, with the Division of Examinations conducting only a fraction of possible inspections in any given year. Human auditors, even when experienced, face fundamental capacity constraints: a single examiner reviewing a hedge fund's monthly return series, trading records, and valuation documentation can assess at most a handful of funds per quarter. This throughput bottleneck means that most hedge funds receive regulatory scrutiny only infrequently, creating long windows during which fraudulent schemes can operate undiscovered.

Beyond capacity, human judgment is subject to well-documented cognitive limitations. The case of Harry Markopolos is instructive. Beginning in 2000, Markopolos repeatedly submitted detailed analyses to the SEC arguing that Madoff's reported returns were statistically implausible, yet the agency failed to act for nearly a decade (**?**). This failure

reflected not only institutional shortcomings but also the difficulty of distinguishing genuine skill from fabrication when evaluating complex, opaque strategies. Hindsight bias, confirmation bias, and anchoring effects further impair the ability of human analysts to identify fraud signals that deviate from established templates.

Traditional statistical methods for fraud detection have provided a valuable foundation but remain insufficient on their own. Benford's law analysis—which tests whether the leading digits of reported returns conform to the expected logarithmic distribution—can identify certain forms of data fabrication but is easily defeated by a knowledgeable fraudster who engineers returns to satisfy digit-frequency tests. Serial correlation analysis, which examines suspicious smoothness in reported return series, has been applied with some success to detect NAV manipulation in hedge funds (Bollen and Pool, 2012; Getmansky et al., 2004), but it captures only one dimension of a potentially multi-faceted fraud. Similarly, forensic ratio analysis and outlier detection based on univariate distributional properties can flag individual anomalies without capturing the complex, multi-dimensional patterns that characterize sophisticated schemes. Dimmock and Gerken (2012) demonstrated that regulatory disclosure data can predict future fraud, yet their logistic regression approach, while interpretable, operates on a limited feature space and does not scale to the volume or variety of data now available.

Artificial intelligence (AI) and machine learning (ML) methods offer four fundamental advantages that address these limitations. First, *scalability*: ML algorithms can process thousands of fund return series, regulatory filings, and alternative data sources simultaneously, enabling surveillance at a scale that human analysts cannot achieve. Second, *pattern recognition*: methods ranging from ensemble classifiers to deep neural networks can detect subtle, nonlinear, and multi-dimensional anomalies that elude univariate statistical tests. For example, a random forest trained on dozens of return-based features—including higher moments, serial correlation coefficients, and distributional shape statistics—can identify suspicious combinations of characteristics that no single test would flag (Bollen and Pool, 2012). Third, *real-time monitoring*: once trained and deployed, ML models can evaluate incoming data continuously, enabling early warning systems that alert regu-

lators and investors to emerging risks before losses compound. Fourth, *multi-modal data integration*: modern AI architectures can fuse structured data (return series, regulatory filings, financial ratios) with unstructured data (news articles, social media sentiment, legal documents) and relational data (counterparty networks, manager affiliation graphs), constructing a richer and more holistic picture of fund behavior than any single data modality permits.

Despite this promise, the application of AI to hedge fund fraud detection remains fragmented, methodologically heterogeneous, and largely disconnected from the operational realities of regulatory enforcement. Existing studies span multiple disciplines—computer science, finance, accounting, and law—employ divergent datasets, evaluation metrics, and fraud definitions, and rarely address the adversarial dynamics inherent in financial fraud, where perpetrators actively adapt their behavior to evade detection. This fragmentation motivates the present survey.

1.3 Survey Scope and Contributions

This paper presents a systematic, qualitative survey of AI-based approaches to hedge fund fraud detection. We synthesize the scattered literature into a coherent analytical framework, identify critical gaps, and propose a concrete research agenda. To the best of our knowledge, no existing survey addresses AI-based fraud detection with a specific focus on the hedge fund context, its unique data challenges, and its distinctive regulatory environment.

Several prior surveys have examined the broader intersection of AI and financial fraud. Ngai et al. (2011) provided an early taxonomy of data mining techniques applied to financial fraud, covering credit card fraud, insurance fraud, and securities fraud but without distinguishing hedge funds from other financial institutions. Abdallah et al. (2016) reviewed fraud detection systems across multiple domains and emphasized the importance of class imbalance, yet their treatment of investment fraud remained cursory. West and Bhattacharya (2016) surveyed intelligent financial fraud detection, focusing primarily on credit card and payment fraud with limited attention to asset management.

Table 1: Comparison of this survey with prior related surveys. Checkmarks (✓) indicate that a topic is substantively addressed; dashes (–) indicate peripheral or absent coverage.

Survey	Hedge Fund Focus	AI/ML Methods	Fraud Taxonomy	Adversarial Robustness	Regulatory Readiness	I
Ngai et al. (2011)	–	✓	–	–	–	
Abdallah et al. (2016)	–	✓	–	–	–	
West and Bhattacharya (2016)	–	✓	–	–	–	
Pourhabibi et al. (2020)	–	✓	–	–	–	
Bao et al. (2020)	–	✓	–	–	–	
Hilal et al. (2022)	–	✓	–	–	–	
Ahmed et al. (2024)	–	✓	–	–	–	
This survey	✓	✓	✓	✓	✓	

Pourhabibi et al. (2020) examined fraud detection using process mining and network analysis, contributing valuable methodological perspectives but not addressing the specific data ecosystem of hedge funds. Bao et al. (2020) focused on detecting financial statement fraud using deep learning, with an accounting rather than investment management orientation. Hilal et al. (2022) offered a comprehensive survey of financial fraud detection methods, covering a broad taxonomy of techniques but treating hedge fund fraud only peripherally. More recently, Ahmed et al. (2024) provided a wide-ranging survey of AI for financial crime detection, and several reviews have examined graph neural networks (GNNs) for fraud detection in financial networks, yet none of these works systematically maps AI detection capabilities to the specific fraud typologies, data structures, and regulatory constraints that characterize the hedge fund industry. Table 1 summarizes the scope of these existing surveys and highlights the gaps that the present work addresses.

We make three principal contributions:

C1: Detection Pipeline Taxonomy. We propose a unified five-stage framework—spanning data ingestion, feature engineering, model selection, explainability, and deployment—that systematically maps hedge fund fraud types to appropriate AI detection methods. This taxonomy provides researchers and practitioners with a structured lens for understanding which methods apply to which fraud scenarios and where methodological gaps remain. No existing survey provides this hedge-fund-specific mapping.

C2: Adversarial and Regulatory Readiness Assessment. We conduct a system-

atic evaluation of how robust current AI detection methods are to adversarial manipulation by sophisticated hedge fund managers, and we assess whether these methods satisfy emerging regulatory requirements, including the European Union Artificial Intelligence Act (EU AI Act) and SEC guidance on the use of predictive analytics. This assessment bridges the gap between the technical ML literature and the practical demands of regulators and compliance professionals. No prior survey has evaluated AI-based fraud detection through this dual lens.

C3: Actionable Research Roadmap. We identify ten concrete open research problems, each differentiated by the specific characteristics of the hedge fund context. For each problem, we suggest methodological approaches, outline evaluation protocols, and discuss feasibility considerations. This roadmap is designed to guide both academic researchers seeking impactful problems and industry practitioners seeking evidence-based solutions.

It is important to note what this survey does *not* attempt. We do not conduct a quantitative meta-analysis of detection performance across studies, as the heterogeneity of datasets, fraud definitions, evaluation protocols, and reporting standards across the existing literature precludes meaningful statistical aggregation. Instead, we adopt a qualitative synthesis approach, critically analyzing the methodological strengths, limitations, and contextual applicability of each body of work. This approach is appropriate given the current state of the field, where standardization of benchmarks and evaluation procedures remains an open challenge that we address explicitly in our research agenda.

1.4 Paper Organization

The remainder of this paper is organized as follows. Section 2 establishes the necessary background, presenting a taxonomy of hedge fund fraud types, describing the data ecosystem available for detection research, and summarizing the regulatory context within which detection systems must operate. Section 3 introduces our detection pipeline framework (Contribution C1), detailing the five stages through which raw data are transformed

into actionable fraud assessments. Section 4 provides a comprehensive qualitative review of AI and ML methods that have been applied—or proposed for application—to hedge fund fraud detection, organized by methodological family and mapped onto our pipeline taxonomy. Section 5 addresses adversarial robustness, regulatory readiness, and ethical considerations (Contribution C2), evaluating the extent to which current methods withstand strategic manipulation and satisfy legal requirements. Section 6 presents our research agenda (Contribution C3), articulating ten open problems with suggested approaches and evaluation criteria. Finally, Section 7 concludes with a synthesis of key findings and implications for researchers, regulators, and practitioners.

2 Background: The Hedge Fund Fraud Landscape

Hedge funds occupy a distinctive position in the financial ecosystem: they manage trillions of dollars in assets, yet operate under substantially lighter disclosure obligations than mutual funds or public companies. This opacity, combined with complex trading strategies and performance-linked compensation structures, creates fertile ground for fraudulent conduct. Before examining how artificial intelligence can detect such fraud (Sections 3 and 4), we must first understand what forms hedge fund fraud takes, which data sources are available for detection, and what regulatory structures govern both the funds and the emerging AI tools applied to their oversight.

2.1 Taxonomy of Hedge Fund Fraud

We organize hedge fund fraud into five categories, ordered roughly from the most frequently studied in the academic literature to the most difficult to detect with currently available data. For each category we provide a definition, a detection difficulty rating on a five-point scale, at least one notable enforcement case, and the observable signals that machine learning models can exploit. Table 2 summarizes the taxonomy.

Table 2: Taxonomy of hedge fund fraud types with detection difficulty ratings and principal observable signals. Difficulty is rated on a scale from 1 (straightforward with available data) to 5 (requires privileged real-time data and advanced methods).

Fraud Type	**Difficulty**	**Key Case**	**Observable Signals**
Performance fabrication	3/5	Madoff (2008)	Serial correlation, Benford violations, implausible Sharpe ratios
Allocation fraud	4/5	Petters (2008)	Cross-account return dispersion, win-rate asymmetry
Strategy misrepresentation	3/5	Platinum Partners (2016)	Style drift, factor exposure shifts, textual inconsistencies
Market manipulation	5/5	SAC Capital (2013)	Order-flow anomalies, network centrality, timing patterns
Regulatory fraud	2/5	Lancer Mgmt. (2003)	Filing inconsistencies, text anomalies, omission detection

2.1.1 Performance Fabrication

Performance fabrication encompasses any deliberate misstatement of investment returns, ranging from outright Ponzi schemes to subtler forms of return smoothing and net asset value (NAV) manipulation. In a Ponzi scheme, new investor capital is used to pay purported returns to existing investors, requiring no actual profitable trading. Return smoothing involves selectively deferring the recognition of losses or spreading gains across periods to produce an artificially stable return series. NAV manipulation inflates the reported value of illiquid positions—such as over-the-counter derivatives, distressed debt, or private placements—for which no reliable market price exists.

The paradigmatic case remains Bernard L. Madoff Investment Securities, which sustained fabricated returns for at least two decades before collapsing in December 2008 with estimated losses of $65 billion (Gregoriou and Lhabitant, 2009). Madoff's reported track record exhibited only seven losing months across a 14-year span, producing a nearly perfect 45-degree equity curve—a statistical near-impossibility for any legitimate trading strategy. **?** famously documented these anomalies years before the scheme was uncovered, noting that the reported Sharpe ratio exceeded plausible bounds for the purported

split-strike conversion strategy.

Detection difficulty: 3/5. Statistical red flags for fabricated returns are well established in the literature. Bollen and Pool (2012) identified several distributional anomalies—including discontinuities at zero in the return distribution, abnormally low serial correlation of losses, and suspiciously consistent positive returns—that collectively flag roughly 8% of funds in the Lipper TASS database as potentially suspicious. Getmansky et al. (2004) developed an econometric model demonstrating that serial correlation in hedge fund returns often arises from the managed pricing of illiquid assets, providing a quantitative baseline against which smoothed returns can be measured. Benford's law, which predicts the frequency distribution of leading digits in naturally occurring numerical data (Benford, 1938; Nigrini, 2012), has also been applied to hedge fund returns: deviations from the expected distribution can signal data fabrication, though the test has limited power for short return histories. While these signals are individually noisy, their combination through machine learning classifiers offers substantially improved discriminatory power—a theme we develop in Section 4.

2.1.2 Allocation Fraud

Allocation fraud occurs when a fund manager systematically directs profitable trades to favored accounts—typically proprietary or co-investment vehicles—while routing losing trades to client accounts. A related variant, cherry-picking, involves delaying the allocation of executed trades until after the daily profit or loss is known, at which point winning trades are assigned to preferred accounts. Securities and Exchange Commission (SEC) enforcement data reveal cases in which favored accounts received 91% profitable trade allocations compared with only 31% for general client accounts, a disparity that cannot arise by chance.

Detection difficulty: 4/5. Allocation fraud is inherently difficult to detect because it requires trade-level data—specifically, the timestamps of order execution and the subsequent assignment of fills to accounts. Such data are rarely available in public databases. Hedge fund return databases (discussed in Section 2.2.1) report only fund-level monthly

returns, which obscure intra-fund allocation patterns entirely. Even with account-level return data, detection requires statistical comparison of return distributions across accounts managed by the same adviser, making cross-account dispersion analysis and win-rate asymmetry the primary observable signals. Network-based approaches that map adviser–account relationships from regulatory filings show promise but remain largely unexplored.

2.1.3 Strategy Misrepresentation

Strategy misrepresentation arises when a fund's actual investment behavior diverges materially from its stated strategy without adequate disclosure. This category includes undisclosed style drift—where a fund marketed as equity-long/short gradually concentrates into illiquid credit positions—as well as leverage misreporting and, increasingly, "AI-washing," the practice of falsely claiming that investment decisions are driven by artificial intelligence or machine learning models when they are not.

Patton et al. (2015) developed change-point detection methods to identify structural breaks in hedge fund risk exposures, showing that style drift often precedes fund failure. Quantitative style analysis using rolling-window regressions against factor benchmarks (Fung and Hsieh, 2001) can detect drift, but it cannot distinguish between legitimate strategy evolution and fraudulent misrepresentation without reference to the fund's disclosure documents. Natural language processing (NLP) applied to Form ADV brochures, offering memoranda, and investor letters can bridge this gap by comparing stated strategy descriptions against quantitative factor exposures.

Detection difficulty: 3/5. The combination of quantitative style analysis and NLP on regulatory filings makes this category amenable to AI-based detection. The principal challenge lies in establishing an appropriate threshold: some degree of style drift is normal and even desirable in dynamic markets, and distinguishing intentional misrepresentation from adaptive portfolio management requires contextual judgment.

2.1.4 Market Manipulation

Market manipulation by hedge funds encompasses front-running of client orders, insider trading (including "shadow trading" on economically related securities to avoid detection), and spoofing—the placement and rapid cancellation of large orders to create a false impression of supply or demand. The 2013 guilty plea by SAC Capital Advisors to insider trading charges, resulting in a $1.8 billion penalty, illustrated the scale at which hedge funds can engage in information-based manipulation (Lewis, 2012).

Detection difficulty: 5/5. Market manipulation is the most difficult fraud category to detect because it requires real-time, tick-level trade and order-book data, which are not available in standard hedge fund databases. Detection methods must incorporate network analysis to identify communication patterns between traders and information sources, temporal analysis of order placement relative to material nonpublic events, and cross-market surveillance to detect shadow trading across related securities. Regulators have begun deploying such systems—the SEC's Market Information Data Analytics System (MIDAS), for instance, ingests and analyzes billions of order and trade records daily—but academic research in this area remains constrained by data access limitations.

2.1.5 Regulatory Fraud

Regulatory fraud involves the submission of materially false or misleading information in mandatory filings. For hedge fund advisers, the primary filings include Form ADV (the uniform registration document required under the Investment Advisers Act of 1940), Form D (Regulation D offering notices), and Form 13F (quarterly holdings reports for institutional investment managers). Fraud in this category ranges from deliberate misstatements—such as understating assets under management (AUM) to avoid registration thresholds—to material omissions, such as failing to disclose disciplinary history or conflicts of interest.

Dimmock and Gerken (2012) demonstrated that information in Form ADV filings has predictive power for subsequent SEC enforcement actions, with prior regulatory sanctions, ownership structures, and custody arrangements serving as significant predictors. Brown et al. (2008) examined the incremental information value of mandatory hedge

fund disclosure, finding that filing data contain fraud-relevant signals that complement return-based statistical tests.

Detection difficulty: 2/5. Regulatory fraud is the most tractable category for AI-based detection because the underlying data are structured (or semi-structured), publicly accessible through the SEC's Electronic Data Gathering, Analysis, and Retrieval (EDGAR) system, and amenable to both traditional text analysis and modern NLP. Cross-referencing filings with external data—for example, verifying reported AUM against fund flows implied by return data—can reveal inconsistencies with high precision.

2.2 The Data Ecosystem for Hedge Fund Fraud Detection

The effectiveness of any fraud detection system is bounded by the quality, coverage, and granularity of its input data. We organize the data landscape into four layers: return data, regulatory filings, alternative data, and synthetic data. Each layer presents distinct opportunities and challenges for AI-based detection.

2.2.1 Return Data

The primary sources of hedge fund return data are commercial databases maintained by Lipper TASS (now Refinitiv), Hedge Fund Research (HFR), BarclayHedge, and Morningstar. The Lipper TASS database, the most widely used in academic research, contains monthly return series for over 7,000 live and defunct funds, along with self-reported fund characteristics such as strategy classification, inception date, management and incentive fees, and lockup provisions.

These databases suffer from three well-documented biases that directly affect fraud detection research. *Survivorship bias* arises because funds that cease reporting—due to liquidation, poor performance, or regulatory action—are removed from the live database. Fung and Hsieh (2009) estimated that survivorship bias overstates average hedge fund returns by approximately 242 basis points per year. *Backfill bias* (or instant history bias) occurs when a fund that begins reporting to a database retroactively submits its prior return history, which tends to be favorably selected; Fung and Hsieh (2009) estimated this

bias at 442 basis points per year. *Selection bias* stems from the voluntary nature of hedge fund reporting: funds with strong track records are more likely to report (for marketing purposes), while distressed or fraudulent funds may stop reporting before detection.

For fraud detection specifically, these biases create a pernicious asymmetry: fraudulent funds that are eventually detected and shut down exit the database, while fraudulent funds that evade detection continue to report. Models trained on survivorship-biased data may therefore underestimate the base rate of fraud and miss the statistical signatures of currently active fraudulent schemes. Researchers must take care to use databases that include "graveyard" (defunct) funds and to model the reporting decision process itself as a potential signal (Agarwal et al., 2011; Aragon, 2007).

2.2.2 Regulatory Filings

Since the passage of the Dodd-Frank Wall Street Reform and Consumer Protection Act in 2010, investment advisers managing $150 million or more in AUM have been required to register with the SEC and file Form ADV. This mandate substantially expanded the universe of hedge funds subject to regulatory disclosure, creating a rich data source that did not exist prior to 2012 for most hedge fund advisers.

The SEC's EDGAR system provides public access to Form ADV (Parts 1 and 2), Form D notices, and Form 13F quarterly holdings. Form ADV Part 1 contains structured data on the adviser's business, ownership, clients, and disciplinary history. Part 2 (the "brochure") is a narrative document describing the adviser's strategies, fee structures, risk factors, and conflicts of interest. Form 13F provides quarterly snapshots of equity holdings for managers with more than $100 million in qualifying securities, enabling holdings-based style analysis.

Parsing these filings presents significant practical challenges. While Form ADV Part 1 is filed electronically in a structured XML format through the Investment Adviser Registration Depository (IARD), Part 2 is a free-text PDF with no standardized structure. Form 13F data, though machine-readable, contain known errors and inconsistencies in security identification. Merging data across filing types and linking to commercial return

databases (e.g., matching TASS fund records to SEC adviser records) requires substantial data engineering effort (Brown et al., 2008; Dimmock and Gerken, 2012). Despite these obstacles, the combination of return data with regulatory filings yields a materially enriched dataset: return-based statistical flags can be cross-validated against filing-derived signals such as auditor changes, custody arrangements, and disciplinary histories.

2.2.3 Alternative Data

The rapid growth of alternative data—non-traditional data sources processed through computational methods—has opened new avenues for fraud detection that extend beyond the return and filing data traditionally available to researchers and regulators. The global alternative data market was estimated at approximately $7.5 billion in 2023, with projections reaching $273 billion by 2032, driven primarily by adoption among investment managers and financial regulators.

Relevant alternative data sources for hedge fund fraud detection include news and social media sentiment, which can flag reputational signals and emerging concerns about specific funds or managers before they appear in regulatory actions. Satellite imagery and geolocation data have been used to independently verify economic claims—for instance, estimating retail foot traffic or commodity inventory levels to cross-check reported fund performance. Web traffic analytics, patent filings, and litigation records provide additional dimensions for triangulating the plausibility of stated strategies and returns.

However, alternative data carry their own risks. Sentiment-derived signals are noisy and susceptible to manipulation through coordinated social media campaigns. The costs of acquiring and processing satellite and geolocation data are substantial, and the relationship between these data and fund-level fraud signals is often indirect. Moreover, the use of alternative data for surveillance raises privacy concerns that intersect with emerging regulatory frameworks, particularly the European Union's AI Act (discussed in Section 2.3.2). Nevertheless, for regulators and fund-of-funds managers with access to these data, they represent a valuable complementary layer that AI systems can integrate through multi-modal learning architectures.

2.2.4 Synthetic Data

A fundamental obstacle in hedge fund fraud detection is the extreme class imbalance inherent in the problem: confirmed fraud cases represent a small fraction of the total fund population, making it difficult to train supervised classifiers with adequate positive-class representation. Synthetic data generation techniques address this imbalance by creating artificial examples of fraudulent fund behavior.

The Synthetic Minority Over-sampling Technique (SMOTE; Chawla et al., 2002) remains the most widely used approach, generating synthetic minority-class samples by interpolating between existing positive examples in feature space. More recent methods employ generative adversarial networks (GANs) and variational autoencoders (VAEs) to produce synthetic return series and feature vectors that capture the distributional properties of known fraudulent funds. These generative approaches can produce more realistic and diverse synthetic samples than interpolation-based methods, but they introduce their own validation challenges: synthetic data must preserve the statistical dependencies and temporal dynamics of real fraud cases without amplifying artifacts of the training data.

Privacy-preserving synthetic data generation is an active area of research with particular relevance for cross-institutional collaboration. Regulators who possess enforcement-labelled datasets cannot share them with academic researchers due to confidentiality constraints. Differentially private generative models could enable the release of synthetic fraud datasets that preserve aggregate statistical properties while protecting the identities of individual funds and managers. Although this approach remains largely aspirational in the hedge fund domain, it has shown promise in adjacent areas of financial crime detection, including anti-money laundering and credit fraud.

2.3 Regulatory Context

The regulatory environment shapes both the data available for fraud detection and the constraints under which AI-based detection systems must operate. We review the principal frameworks in the United States and European Union, followed by the emerging field of supervisory technology (SupTech).

2.3.1 United States

The US regulatory framework for hedge fund oversight underwent a structural transformation following the 2008 financial crisis and the Madoff scandal. Title IV of the Dodd-Frank Act eliminated the "private adviser exemption" that had previously allowed most hedge fund advisers to avoid SEC registration, requiring advisers with AUM of $150 million or more to register and file Form ADV. This single regulatory change dramatically expanded the universe of funds subject to systematic oversight and created the filing data that now underpin many detection approaches (Brown et al., 2008).

The SEC has invested substantially in computational enforcement capabilities. The Division of Economic and Risk Analysis (DERA), established in 2009, provides quantitative analysis to support enforcement investigations and rulemaking. MIDAS, operational since 2013, collects and analyzes data from all equity exchanges and off-exchange venues, processing approximately one billion records per day to detect market manipulation patterns. The Center for Risk and Quantitative Analytics (CRQA), housed within the Office of Compliance Inspections and Examinations, maintains databases of trading patterns derived from past enforcement actions and uses these to prioritize future examinations (**?**).

The SEC's Whistleblower Program, established under Dodd-Frank Section 922 in response to the failure to act on **?**'s (**?**) repeated warnings about Madoff, has become a significant source of enforcement leads. Since its inception, the program has awarded over $1.5 billion to whistleblowers and generated thousands of tips that complement algorithmic detection. The interaction between human intelligence (whistleblower tips) and machine intelligence (quantitative screening) represents an underexplored hybrid detection paradigm.

2.3.2 European Union

In the European Union, the Alternative Investment Fund Managers Directive (AIFMD), adopted in 2011, established a harmonized regulatory framework for managers of alternative investment funds, including hedge funds. AIFMD imposes reporting obliga-

tions, leverage limits, and investor disclosure requirements that generate structured data analogous—though not identical—to the SEC's Form ADV regime.

More directly relevant to AI-based fraud detection is the EU Artificial Intelligence Act (Regulation 2024/1689), which entered into force in August 2024 (European Parliament and Council of the European Union, 2024). The AI Act classifies AI systems used for "creditworthiness assessment" and "fraud detection in financial services" as high-risk, subjecting them to mandatory requirements including risk management systems, data governance standards, technical documentation, human oversight provisions, and transparency obligations. Notably, the Act requires that high-risk AI systems provide outputs that are "sufficiently transparent to enable deployers to interpret the system's output and use it appropriately" (Art. 13), which has direct implications for model selection in fraud detection: opaque models such as deep neural networks may require post-hoc explainability methods (e.g., SHAP, LIME) to satisfy regulatory requirements, while inherently interpretable models such as logistic regression or decision trees may be preferred despite potentially lower predictive performance.

The tension between predictive accuracy and explainability is not merely academic. A detection system that identifies suspicious funds but cannot articulate the basis for its suspicion is of limited use to regulators who must justify enforcement actions in administrative proceedings and courts. The EU AI Act codifies this intuition into law, and similar requirements are likely to emerge in other jurisdictions. We return to this tension in Section 6.

2.3.3 Supervisory Technology

Supervisory technology (SupTech) refers to the use of advanced analytics and AI by financial regulators and central banks to enhance their oversight capabilities (Financial Stability Board, 2017; **?**). The adoption of SupTech for hedge fund oversight represents a shift from reactive enforcement—investigating fraud after losses have materialized—to proactive surveillance that aims to detect anomalies before they escalate.

Several regulatory agencies have reported SupTech initiatives relevant to hedge fund

fraud detection. The SEC's DERA and CRQA units, discussed above, represent early examples. The Bank of England, the Monetary Authority of Singapore, and the De Nederlandsche Bank have all piloted machine learning systems for anomaly detection in financial reporting data. These systems typically operate by establishing baseline behavioral profiles for regulated entities and flagging deviations that exceed statistical thresholds, an approach conceptually similar to the return-based statistical tests pioneered by Bollen and Pool (2012) but applied at institutional scale across multiple firms simultaneously.

Despite these advances, SupTech adoption faces significant challenges. Regulators must manage the risk of false positives, which consume scarce examination resources, while simultaneously ensuring that sophisticated fraudsters cannot reverse-engineer detection criteria. The shortage of personnel with combined expertise in financial regulation and machine learning constrains implementation. Cross-border coordination remains limited: a fund registered in the Cayman Islands, managed from New York, with European investors and Asian trading venues may fall within the jurisdiction of multiple regulators, none of whom possess a complete picture. AI-based detection systems that operate on fragmented, jurisdiction-specific data inevitably produce a partial view of potentially global fraudulent schemes.

This section has established the conceptual and empirical foundations for the remainder of the paper. The fraud taxonomy (Section 2.1) identifies what must be detected; the data ecosystem (Section 2.2) characterizes the inputs available to detection systems; and the regulatory context (Section 2.3) defines the institutional constraints under which these systems operate. Section 3 builds on this foundation by specifying the technical architecture of an end-to-end AI-based detection pipeline.

3 A Unified Detection Pipeline Framework

The preceding section catalogued the fraud types that afflict the hedge fund industry, the data sources available for detection, and the regulatory environment within which surveillance systems must operate. We now synthesize these elements into a unified, end-to-end

detection pipeline—the paper's primary organizational contribution (Contribution C1). While individual components of this pipeline have been studied in isolation—serial correlation analysis for return smoothing (Getmansky et al., 2004), Benford's law tests for data fabrication (Nigrini, 2012), logistic regression on regulatory filings for fraud prediction (Dimmock and Gerken, 2012)—no prior work has assembled them into a coherent architectural framework tailored specifically to the hedge fund context. The five-stage pipeline presented here maps the fraud taxonomy of Section 2.1 to concrete detection stages, identifies which AI and machine learning methods are best suited to each stage, and exposes the interfaces and feedback mechanisms through which stages interact. This framework serves three purposes: it provides researchers with a structured lens for positioning new contributions within the detection workflow; it offers practitioners an engineering blueprint for building operational surveillance systems; and it reveals the methodological gaps that motivate the research agenda of Section 6.

3.1 Pipeline Overview

The detection pipeline comprises five sequential stages, each transforming the output of its predecessor into progressively more refined and actionable assessments:

1. **Data Ingestion and Integration** (Section 3.2): Multi-source data collection, temporal alignment, entity resolution, and quality assurance across heterogeneous data streams.

2. **Feature Engineering** (Section 3.3): Extraction of domain-specific features organized into five families—statistical, Benford's law, textual, network-relational, and temporal—that translate raw data into representations amenable to machine learning.

3. **Model Selection and Training** (Section 3.4): Selection and training of AI methods, organized into six method families, matched to specific fraud types and data characteristics.

4. **Explainability and Interpretation** (Section 3.5): Post-hoc and intrinsic explanation of model outputs to satisfy regulatory transparency requirements and support human decision-making.

5. **Deployment and Monitoring** (Section 3.6): Operationalization of trained models within production surveillance environments, including concept drift detection, alert prioritization, and human-in-the-loop feedback.

Critically, the pipeline is not strictly unidirectional. A feedback loop from the deployment stage back to feature engineering and model training enables the system to incorporate investigator judgments, adapt to evolving fraud patterns, and recalibrate as new data sources become available. Figure 1 illustrates the overall architecture, including data flow arrows between stages, feedback loops from deployment to earlier stages, and annotations indicating which fraud types from Table 2 are primarily addressed at each stage.

3.2 Stage 1: Data Ingestion and Integration

The first pipeline stage addresses the fundamental challenge of assembling a coherent analytical dataset from data sources that differ in structure, frequency, reliability, and provenance. As detailed in Section 2.2, the hedge fund data ecosystem spans at least four layers: return data from commercial databases (monthly, numerical), regulatory filings from SEC EDGAR and equivalent repositories (quarterly or annually, textual and semi-structured), alternative data from news, social media, and satellite providers (continuous, heterogeneous), and synthetic data generated to address class imbalance (on-demand, model-derived). Fusing these sources into a unified representation suitable for downstream feature extraction requires addressing four distinct sub-problems.

Temporal alignment. Each data source operates on a different reporting cadence. Return data arrive monthly with a typical lag of 30 to 60 days. Form ADV filings are updated annually, with material amendments filed on an ad hoc basis. Form 13F holdings

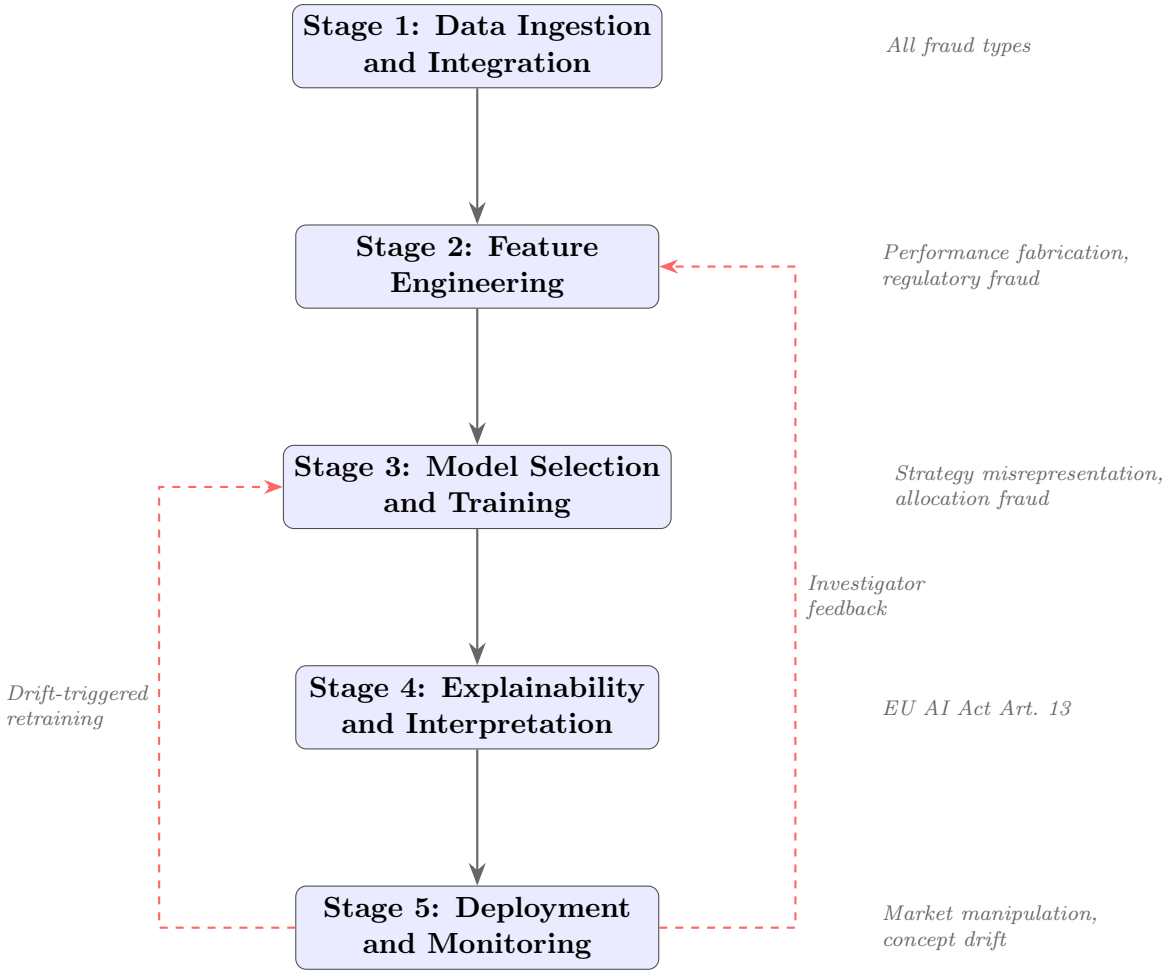

Figure 1: Architecture of the five-stage detection pipeline. Solid arrows indicate the forward data flow; dashed arrows represent feedback loops from deployment to earlier stages. Annotations on the right indicate which fraud types (Table 2) and regulatory considerations are primarily addressed at each stage.

are disclosed quarterly with a 45-day delay. News and social media data arrive continuously but at irregular intervals. Aligning these streams requires careful interpolation and aggregation decisions: monthly returns can be aligned to quarter-end dates to match filing data, but this discards intra-quarter dynamics that may carry fraud signals. Alternative data must be aggregated into windows that match the periodicity of the lowest-frequency source, or multi-resolution architectures must be employed that process each stream at its native frequency before fusion. The choice of alignment strategy directly affects which temporal patterns—such as the end-of-quarter return spikes that Bollen and Pool (2012) identified as potential manipulation signals—remain visible to downstream models.

Data quality and bias correction. Hedge fund data suffer from well-documented biases that, if left uncorrected, propagate through the entire pipeline. Survivorship bias, estimated by Fung and Hsieh (2009) at approximately 242 basis points per year in return overstatement, arises because defunct funds exit live databases. Backfill bias, estimated at 442 basis points per year, results from the retroactive inclusion of favorable pre-reporting return histories. Self-selection bias reflects the voluntary nature of database reporting: funds with poor or suspicious performance may simply stop reporting rather than reveal deteriorating results (Agarwal et al., 2011; Aiken et al., 2013). For fraud detection specifically, these biases create a pernicious asymmetry—fraudulent funds that are eventually detected and shut down enter the graveyard section of databases, while successful frauds that continue operating remain in the live section, potentially contaminating the "clean" training class. Correction procedures include restricting analysis to post-reporting-inception returns (to address backfill bias), requiring the use of databases that maintain graveyard records (to address survivorship bias), and modeling the reporting decision process itself as an informative signal (Agarwal et al., 2011).

Entity resolution. A single hedge fund may appear under different names, identifiers, and organizational structures across data sources. The Lipper TASS database assigns its own fund identifiers, which do not correspond to HFR's classification scheme, the SEC's Central Registration Depository (CRD) numbers, or the Legal Entity Identifiers (LEIs) used in European regulatory filings. A fund manager who launches a new vehicle after a prior fund's failure may deliberately obscure the connection between the old and new entities. Entity resolution—the task of linking records that refer to the same real-world fund or manager—is therefore both a data engineering challenge and a fraud-relevant signal: the inability to link a fund to its predecessors may itself indicate an attempt to escape reputational consequences. Approximate string matching on fund and manager names, combined with shared-attribute clustering (common addresses, auditors, administrators, or prime brokers), provides a practical starting point. Graph-based entity resolution methods, which propagate identity evidence through networks of shared relationships, offer more robust linking at the cost of greater computational complexity. The Dodd-

Frank Act's mandate of Form ADV filing since 2010 has substantially improved entity resolution for US-registered advisers by providing a stable regulatory identifier (the CRD number) that can serve as a linkage key across databases (Brown et al., 2008).

Multi-source fusion architecture. The ingestion stage must produce a unified fund-level record that integrates all available data modalities. We distinguish two architectural approaches. In *early fusion*, raw data from all sources are merged into a single feature matrix before any feature extraction occurs; this approach is simple but requires all sources to be aligned to a common temporal grid and forces a single representation for fundamentally different data types. In *late fusion*, each data source is processed through a modality-specific feature extraction pipeline (described in Section 3.3), and the resulting feature vectors are concatenated or combined through learned attention weights at the model input stage. Late fusion preserves the native structure of each data source and accommodates missing modalities gracefully—a fund that does not file Form 13F because it falls below the reporting threshold can still be analyzed using return and filing data—but it requires careful design of the fusion mechanism to ensure that cross-modal interactions (such as inconsistencies between stated strategy and realized factor exposures) are captured rather than lost.

3.3 Stage 2: Feature Engineering

The feature engineering stage transforms the integrated dataset from Stage 1 into a structured representation that encodes the statistical, linguistic, relational, and temporal signatures of hedge fund fraud. We organize features into five families, each targeting different fraud types and exploiting different data modalities. The full mathematical specifications are provided in Section B; here we describe the intuition, key examples, and fraud-detection rationale for each family.

3.3.1 Statistical Features

Statistical features operate on the monthly return series $\{r_t\}_{t=1}^{T}$ and capture distributional anomalies that arise when returns are fabricated, smoothed, or otherwise manipulated.

Serial correlation. The first-order autocorrelation coefficient $\rho_1 = \text{Corr}(r_t, r_{t-1})$ serves as a proxy for return smoothing and illiquidity-induced stale pricing. Getmansky et al. (2004) developed an econometric model demonstrating that managed pricing of illiquid assets introduces predictable serial correlation into reported returns, with typical ρ_1 values of 0.3 to 0.5 for funds holding illiquid positions. Abnormally high serial correlation, particularly when the fund claims to hold liquid assets, signals potential NAV manipulation. Higher-order autocorrelations ρ_2, ρ_3 and the Ljung-Box Q-statistic provide complementary measures of return predictability that legitimate returns should not exhibit.

Distributional discontinuity. Bollen and Pool (2012) identified a distinctive "kink" at zero in the empirical return distribution of suspicious funds: an excess of small positive returns and a deficit of small negative returns, consistent with managers selectively reclassifying marginal losses as small gains. This discontinuity can be quantified by comparing the density of returns in a narrow band above zero to the density in a symmetric band below zero, or by testing for a structural break in the return histogram at zero using kernel density estimation.

Higher moments and risk-adjusted performance. The Sharpe ratio $S = \bar{r}/\sigma_r$ provides a baseline measure of risk-adjusted performance; implausibly high Sharpe ratios, as in the Madoff case, signal fabrication. Skewness and excess kurtosis capture asymmetry and tail behavior: legitimate hedge fund returns typically exhibit negative skewness and positive excess kurtosis, reflecting exposure to tail risk, while fabricated returns often display near-zero skewness and low kurtosis (Lo, 2001). Maximum drawdown—the largest peak-to-trough decline in the cumulative return series—provides an additional measure of downside exposure that is difficult to fabricate convincingly over long horizons.

Long-memory detection. The Hurst exponent H, estimated via rescaled range (R/S) analysis or detrended fluctuation analysis (DFA), measures the degree of long-range de-

pendence in a return series. A Hurst exponent significantly above 0.5 indicates persistent serial dependence that may arise from return smoothing or managed pricing. Combined with GARCH-model residual analysis, which captures volatility clustering patterns, the Hurst exponent adds a fractal dimension to the characterization of return dynamics.

3.3.2 Benford's Law Features

Benford's law, first observed empirically by Benford (1938) and formalized by Nigrini (2012) for forensic applications, predicts that the leading digit d of numbers drawn from many naturally occurring distributions follows the probability $P(d) = \log_{10}(1 + 1/d)$ for $d \in \{1, 2, \ldots, 9\}$. Deviations from this expected distribution in reported financial data can signal fabrication, since human-generated or algorithmically engineered numbers often fail to reproduce the logarithmic digit pattern.

First-digit test. The observed frequency of each leading digit in a fund's return series (or reported net asset values) is compared to the Benford distribution using a chi-squared goodness-of-fit test:

$$\chi^2 = \sum_{d=1}^{9} \frac{(O_d - E_d)^2}{E_d}, \tag{1}$$

where O_d is the observed count and $E_d = N \cdot \log_{10}(1 + 1/d)$ is the expected count for N total observations. Significant departures indicate potential data manipulation, though the test has limited statistical power for short return histories typical of hedge funds (e.g., fewer than 60 monthly observations).

Second-digit and summation tests. The second-digit test examines digits in the tens place and is often more sensitive to subtle manipulation than the first-digit test, because fabricators who are aware of Benford's law may engineer first digits to conform while neglecting second digits (Nigrini, 2012). The summation test compares the sum of values sharing each leading digit to the theoretically expected proportion; it is particularly effective for detecting round-number manipulation in reported amounts.

Multi-dimensional Benford feature space. For machine learning applications, the nine first-digit frequencies, ten second-digit frequencies, and two test statistics (chi-squared and Kolmogorov-Smirnov) can be concatenated into a 21-dimensional feature vector that

captures the overall conformity of a fund's reported data to Benford expectations. This representation enables ML models to detect complex patterns of digit manipulation that individual hypothesis tests would miss—for example, a fund whose first digits conform to Benford's law but whose second digits are anomalously uniform.

3.3.3 Textual Features from Regulatory Filings

Regulatory filings, particularly the narrative sections of Form ADV Part 2 brochures and investor letters, contain linguistic signals that complement quantitative return analysis. Natural language processing (NLP) methods extract features that capture the complexity, sentiment, consistency, and evolution of a fund's disclosures.

Filing complexity and readability. The Gunning Fog index, defined as $\text{Fog} = 0.4 \times (\text{ASL} + \text{PHW})$ where ASL is average sentence length and PHW is the percentage of hard words (three or more syllables), measures the readability of filing text. Total word count, sentence count, and type-token ratio provide complementary measures. Prior research in accounting fraud has demonstrated that firms engaged in misconduct tend to produce filings of greater linguistic complexity, possibly to obscure material information.

Sentiment analysis. Domain-specific language models, in particular FinBERT (Araci, 2019)—a BERT model (Devlin et al., 2019) fine-tuned on financial text—enable sentiment scoring that captures the tone of fund communications. Aggregate sentiment scores, sentiment volatility (the standard deviation of sentence-level scores across a filing), and the proportion of hedging language ("may," "could," "potential") provide features that can signal managerial uncertainty or deliberate obfuscation. SEC-BERT (Loukas et al., 2022), pre-trained specifically on EDGAR filings, offers further domain adaptation for regulatory text.

Boilerplate deviation and temporal drift. The cosine similarity between a fund's filing text and a template constructed from the mean term-frequency representation of peer filings measures the degree of boilerplate language. Funds that deviate substantially from peer templates—either through unusually vague language or through anomalously specific disclosures—may warrant scrutiny. Critically, the *temporal trajectory* of readability and

boilerplate deviation can be more informative than any single cross-sectional measurement: deteriorating readability over successive filings, increasing use of hedging language, or growing divergence from prior filings may indicate evolving concealment behavior as fraud progresses.

Topic modeling and strategy description analysis. Latent Dirichlet Allocation (LDA) or neural topic models applied to the strategy description sections of Form ADV brochures can identify topic drift over time—shifts in the language used to describe a fund's investment approach that may correspond to undisclosed strategy changes. Comparing the topic distribution of a fund's textual strategy description against its quantitative factor exposures (derived from return-based style analysis) enables cross-modal consistency checking: a fund that describes an equity long/short strategy but whose returns load on credit spread and volatility factors exhibits a text-quant inconsistency that merits investigation.

3.3.4 Network and Relational Features

Hedge funds do not operate in isolation. Each fund is embedded in a network of relationships with service providers (auditors, administrators, custodians, prime brokers), other funds (through co-investment, shared managers, or common limited partners), and regulatory entities. Features derived from these network structures capture "guilt by association" patterns and organizational risk indicators that return-based and textual features cannot detect.

Fund–service-provider graphs. The bipartite graph linking funds to their auditors, administrators, and custodians encodes critical operational risk information. Small, non-Big-Four audit firms are over-represented among funds that subsequently face enforcement actions (Brown et al., 2008). A fund that changes its auditor from a reputable firm to a smaller, less established firm—particularly if the change coincides with other risk signals—exhibits a pattern associated with weakening oversight. Node-level features extracted from this graph include the auditor's client count, the auditor's historical association with sanctioned funds, and the frequency of auditor changes.

Manager history networks. A graph connecting managers to the funds they have operated, including defunct funds, captures the phenomenon of "serial offenders"—managers who launch new funds after prior vehicles have failed or been sanctioned. Dimmock and Gerken (2012) demonstrated that prior regulatory sanctions are a significant predictor of future fraud. Network features such as the number of prior fund closures, the time since the most recent closure, and the degree of overlap in service providers between successive funds encode this managerial track record.

Co-investment and capital flow networks. Funds sharing common investors or exhibiting correlated capital flows may be connected through structures—such as fund-of-funds arrangements or informal referral networks—that facilitate the propagation of fraudulent schemes. Centrality measures computed on these networks provide fund-level features: a fund with unusually high betweenness centrality occupies a bridging position between otherwise disconnected investor communities, a structural signature associated with Ponzi schemes that depend on continuous capital inflows from diverse sources. Degree centrality, eigenvector centrality, and clustering coefficients offer complementary perspectives on a fund's structural role in the capital flow network.

3.3.5 Temporal Features

Time-series dynamics carry fraud-relevant information beyond what static distributional features capture. Temporal features model the evolution of fund behavior and detect transitions that may signal the onset, escalation, or concealment of fraudulent activity.

Regime detection. Hidden Markov Models (HMMs) fitted to a fund's return series can identify latent regime states—for example, a "normal" regime and a "manipulated" regime characterized by different mean, variance, and autocorrelation parameters. The estimated transition probabilities between regimes, the posterior probability of occupying each regime at each time point, and the timing of regime transitions relative to market events provide features that can distinguish legitimate strategy adaptation from suspicious behavioral shifts.

Change-point detection. Patton et al. (2015) demonstrated that structural breaks

in hedge fund risk exposures frequently precede fund failure. Bayesian change-point algorithms, such as the Bayesian Online Change Point Detection (BOCPD) method, detect abrupt shifts in the parameters of a fund's return distribution. The number of detected change points, their temporal spacing, and the magnitude of parameter changes at each transition serve as features. A fund exhibiting frequent, large-magnitude change points that do not correspond to identifiable market events warrants closer examination.

Calendar effects and end-of-period manipulation. Systematic patterns in return timing—such as consistently higher returns in December relative to other months, or abnormally positive returns on the last trading day of each quarter—can signal end-of-period NAV manipulation. These calendar features are computed as the coefficients of month-of-year and day-of-quarter dummy variables in a regression framework, or as the deviation of period-end returns from the fund's baseline return distribution.

Momentum and reversal patterns. The autocorrelation structure of returns at multiple lags captures momentum (positive autocorrelation at short horizons) and reversal (negative autocorrelation at longer horizons) dynamics. While legitimate strategies exhibit characteristic momentum-reversal patterns that reflect their investment style, fabricated returns often display artificially smooth momentum without the mean-reversion that economic fundamentals would impose.

3.4 Stage 3: Model Selection and Training

The model selection stage matches AI and machine learning methods to the specific characteristics of the hedge fund fraud detection problem: extreme class imbalance, small sample sizes, heterogeneous fraud types, multi-modal input features, and adversarial dynamics. We organize available methods into six families, assessing each in terms of its suitability for the hedge fund context rather than abstract performance benchmarks. A detailed comparative review of specific studies within each family is deferred to Section 4; here we describe the method families, their strengths and limitations, and their mapping to fraud types.

3.4.1 Supervised Classification: Classical Machine Learning

Classical supervised methods treat fraud detection as a binary or multi-class classification problem, learning a mapping from feature vectors to fraud labels using labeled training data.

Logistic regression provides the interpretable baseline. Dimmock and Gerken (2012) applied logistic regression to Form ADV features and achieved an area under the receiver operating characteristic curve (AUC) of approximately 0.65 to 0.70 for predicting future SEC enforcement actions. The model's transparency—each coefficient directly quantifies the log-odds contribution of the corresponding feature—makes it well suited to regulatory settings where decisions must be justified. Its principal limitation is the assumption of linear feature-response relationships, which cannot capture the complex, nonlinear interactions among fraud indicators.

Support vector machines (SVMs) construct maximum-margin hyperplanes in feature space and handle moderate-dimensional problems effectively (Cortes and Vapnik, 1995). The one-class SVM variant, trained only on "normal" fund returns, provides a semi-supervised anomaly detection capability that circumvents the need for labeled fraud examples. SVMs perform well on small datasets—a relevant advantage given the limited number of confirmed hedge fund fraud cases—but scale poorly to very high-dimensional feature spaces and provide limited native interpretability.

Random forests (Breiman, 2001) aggregate predictions from hundreds of decision trees, each trained on a bootstrapped subsample with random feature selection. They handle high-dimensional, mixed-type feature spaces naturally, are robust to outliers and missing values, and provide feature importance scores through permutation importance or mean decrease in impurity. In the hedge fund context, random forests can ingest the full concatenated feature vector spanning statistical, Benford, textual, network, and temporal features without requiring dimensionality reduction.

Gradient boosting methods—XGBoost (Chen and Guestrin, 2016), LightGBM (Ke et al., 2017), and CatBoost (Prokhorenkova et al., 2018)—represent the current state of the art for tabular classification tasks and dominate production fraud detection systems

across the financial industry. They construct ensembles of shallow decision trees sequentially, with each tree correcting the residual errors of its predecessors. CatBoost handles categorical features natively, which is useful for encoding strategy classifications, auditor identities, and jurisdiction codes without one-hot encoding. Ensemble stacking, which combines gradient boosting with other model families (e.g., random forests and neural networks) through a meta-learner, has achieved the highest reported detection performance in financial fraud settings, with F_1 scores approaching 0.88 in some studies (Hilal et al., 2022).

3.4.2 Supervised Classification: Deep Learning

Deep learning architectures offer the capacity to learn hierarchical, nonlinear representations directly from raw or minimally processed data, but their application to hedge fund fraud detection faces distinctive challenges.

Long Short-Term Memory (LSTM) networks (Hochreiter and Schmidhuber, 1997), a recurrent architecture designed to capture long-range dependencies in sequential data, are a natural fit for modeling monthly return time series. An LSTM can learn temporal patterns—such as the gradual onset of return smoothing or the sudden transition from legitimate to fabricated reporting—that static feature vectors cannot represent. However, hedge fund return series are extremely short by deep learning standards: 60 to 120 monthly observations per fund, compared to the thousands or millions of time steps available in speech, text, or high-frequency financial data. This data scarcity limits the effective depth and complexity of LSTM architectures and increases the risk of overfitting.

Convolutional neural networks (CNNs), originally developed for image recognition (LeCun et al., 2015), can be adapted to time-series analysis by converting return sequences into two-dimensional representations—for example, recurrence plots or Gramian angular fields—and applying standard convolutional filters. This approach enables the detection of visual patterns in return dynamics, such as the characteristic "smooth upward ramp" of a Ponzi scheme, that are difficult to capture with hand-crafted features. The representational transformation adds a pre-processing step but leverages the powerful

pattern-matching capabilities of convolutional architectures.

Transformer architectures (Vaswani et al., 2017), which use self-attention mechanisms to model dependencies between all pairs of positions in a sequence, can handle the long-range dependencies in sparse monthly time series more effectively than LSTMs when augmented with positional encodings that account for irregular temporal spacing. Temporal attention weights are themselves interpretable: they reveal which historical periods the model considers most relevant to the current fraud assessment, providing a form of built-in explainability. The principal challenge for all deep learning methods in this domain is data efficiency: with only 10,000 to 15,000 funds and fewer than 100 confirmed fraud cases, the parameter-to-data ratio is unfavorable, and regularization, pre-training, and transfer learning strategies are essential to prevent overfitting.

3.4.3 Unsupervised Anomaly Detection

Unsupervised methods detect fraud by identifying funds whose behavior deviates substantially from the learned distribution of "normal" behavior, without requiring labeled fraud examples. This characteristic is particularly valuable in the hedge fund context, where confirmed fraud labels are scarce and potentially biased toward historically detected fraud types (Chandola et al., 2009).

Isolation Forest (Liu et al., 2008) isolates anomalies by recursively partitioning the feature space with random splits; anomalous observations, which occupy sparse regions of the feature space, require fewer splits to isolate and therefore receive higher anomaly scores. The method is computationally efficient, scales well to high-dimensional feature spaces, and requires no distributional assumptions. It is effective for detecting global anomalies—funds that are unusual relative to the entire population—but may miss local anomalies that are unusual only within their strategy peer group.

Local Outlier Factor (LOF) (Breunig et al., 2000) addresses this limitation by measuring the local density deviation of each point relative to its k-nearest neighbors. A fund that appears normal in the global feature space but has anomalously low density relative to funds pursuing similar strategies receives a high LOF score. This local perspective is

critical for hedge fund detection, where strategy-specific norms differ substantially: the return characteristics that are normal for a global macro fund would be anomalous for a statistical arbitrage fund.

Deep autoencoders learn compressed representations of normal fund behavior and flag funds whose reconstruction error—the discrepancy between the original input and the autoencoder's reconstruction—exceeds a threshold (Chalapathy and Chawla, 2019). The latent representation captures the essential statistical regularities of legitimate returns, and deviations in the reconstruction reveal dimensions along which a fund's behavior is anomalous. Variational autoencoders (VAEs) extend this framework with a probabilistic latent space, enabling the computation of explicit likelihood scores for each fund.

Density-based clustering. DBSCAN (Ester et al., 1996) identifies clusters of funds with similar characteristics and labels funds that do not belong to any cluster as noise points—potential anomalies. In the fraud detection context, a fund that cannot be assigned to any strategy peer group may be misrepresenting its investment approach or pursuing an undisclosed strategy, warranting further investigation.

3.4.4 Natural Language Processing and Text Mining

NLP methods extract fraud-relevant signals from the textual data sources identified in Section 3.3.3, enabling detection capabilities that purely quantitative approaches cannot provide.

The evolution of NLP methods in financial applications has followed the broader trajectory of the field: from bag-of-words representations and term frequency–inverse document frequency (TF-IDF) weighting, through distributed word embeddings (word2vec, GloVe), to contextualized language models based on the transformer architecture. Each generation has expanded the range of textual signals that can be captured. Bag-of-words methods can identify the presence of specific fraud-associated terms but miss semantic context; word embeddings capture semantic similarity but not the sequential structure of sentences; transformer-based models capture both semantics and context, enabling nuanced understanding of hedge fund communications.

FinBERT (Araci, 2019), fine-tuned on financial news and analyst reports, provides domain-adapted sentiment analysis that outperforms general-purpose sentiment tools on financial text. Applied to Form ADV brochures, investor letters, and marketing materials, FinBERT can score the overall tone of fund communications and detect sentiment shifts that precede fraud revelations. *SEC-BERT* (Loukas et al., 2022), pre-trained on the full corpus of EDGAR filings, offers further specialization for regulatory documents, capturing the distinctive vocabulary and rhetorical patterns of SEC filings that general financial language models may miss.

Key NLP applications for hedge fund fraud detection include detecting vague or evasive language in strategy descriptions, identifying inconsistencies between successive filings by the same fund, measuring the divergence between a fund's textual strategy description and its quantitative return characteristics, and flagging unusual linguistic patterns in prospectuses and offering memoranda. The combination of NLP-derived features with return-based statistical features creates a multi-modal detection capability that is more robust to evasion than either modality alone: a fraudster who engineers returns to satisfy statistical tests must also maintain textual consistency across filings, substantially raising the complexity of concealment.

3.4.5 Graph Neural Networks

Graph neural networks (GNNs) operate on the relational structures described in Section 3.3.4, propagating information through the fund-entity network to produce node-level embeddings that capture both a fund's own features and the characteristics of its neighbors (Pourhabibi et al., 2020).

Graph Convolutional Networks (GCNs) (Kipf and Welling, 2017) aggregate neighborhood information through spectral graph convolutions, producing embeddings that encode the local network structure around each fund node. A fund connected to a recently sanctioned auditor, a manager with a history of fund failures, and a custodian with weak controls would receive an embedding that reflects these risk-associated connections, even if the fund's own return-based features appear benign.

Graph Attention Networks (GATs) (Veličković et al., 2018) extend GCNs by learning attention weights that determine how much influence each neighbor exerts on a node's embedding. This attention mechanism is particularly useful for heterogeneous fund-entity graphs, where the relevance of different relationship types (auditor vs. administrator vs. co-investor) to fraud risk varies and should be learned from data rather than specified a priori.

GraphSAGE (Hamilton et al., 2017) enables inductive learning on unseen nodes by learning a neighborhood aggregation function rather than node-specific embeddings. This property is essential for hedge fund detection, where new funds continuously enter the market and must be assessed without retraining the full model. A newly launched fund can be evaluated by sampling and aggregating its neighborhood in the fund-entity graph, leveraging the model's learned understanding of which relational patterns are associated with elevated fraud risk.

Temporal knowledge graphs extend static GNN architectures by incorporating time-stamped edges that capture the evolution of fund-entity relationships. A fund that severs its relationship with a reputable auditor and simultaneously establishes connections with service providers associated with prior fraud cases exhibits a temporal relational pattern—a "flight from oversight"—that static network features cannot capture but temporal GNNs can model explicitly.

The key advantage of GNN-based approaches for hedge fund fraud detection is their ability to capture "guilt by association" patterns: funds that are individually inconspicuous but are embedded in suspicious relational structures can be identified through network propagation. The principal limitation is data availability: constructing comprehensive fund-entity graphs requires linking data across multiple sources, a process that depends on the entity resolution capabilities of Stage 1.

3.4.6 Generative and Synthetic Methods

Generative models serve two distinct roles in the detection pipeline: as anomaly detectors that learn the distribution of normal fund behavior, and as data augmentation tools that

generate synthetic fraud examples to address class imbalance.

Generative adversarial networks for anomaly detection. GAN-based anomaly detection methods, including BiGAN and AnoGAN (Goodfellow et al., 2014), train a generator to produce realistic return series that resemble normal fund behavior. At inference time, a fund whose returns cannot be well reconstructed by the generator—measured by the reconstruction error in the latent space—is flagged as anomalous. The adversarial training process forces the generator to capture the full complexity of normal return distributions, including non-Gaussian features and temporal dependencies, producing a more expressive normality model than parametric distributional assumptions.

Synthetic data generation for class imbalance. Beyond SMOTE (Chawla et al., 2002), which generates synthetic minority-class samples through linear interpolation in feature space, conditional GANs and variational autoencoders (Kingma and Welling, 2014) can generate synthetic fraud examples that preserve the statistical dependencies and temporal dynamics of real fraud cases. Conditional generation is essential in the hedge fund context because different fraud types produce qualitatively different statistical signatures—synthetic Ponzi-scheme returns should exhibit different distributional properties than synthetic NAV-manipulation returns. The Wasserstein GAN variant, which optimizes the earth-mover distance between generated and real distributions, has shown particular promise for financial time-series generation. Validation of synthetic data quality is critical: generated samples must pass domain-specific plausibility checks (e.g., realistic return magnitudes, appropriate serial correlation structure, sensible Sharpe ratios) before being used to augment training sets.

3.5 Stage 4: Explainability and Interpretation

A detection system that flags a fund as suspicious but cannot articulate the basis for its assessment is of limited operational value. Regulators who must justify examination priorities, compliance officers who must escalate findings to senior management, and enforcement attorneys who must present evidence in administrative proceedings all require explanations that translate model outputs into actionable intelligence. The EU AI Act's

Article 13 transparency requirement (discussed in Section 2.3.2) codifies this need into law for high-risk AI systems (European Parliament and Council of the European Union, 2024). The explainability stage addresses this requirement through three complementary approaches.

SHAP values. SHapley Additive exPlanations (Lundberg and Lee, 2017) decompose each prediction into the additive contribution of each input feature, grounded in the cooperative game theory concept of Shapley values. For a fund flagged by a gradient-boosted ensemble, SHAP can identify that, for example, 35% of the fraud score is attributable to abnormally high first-order serial correlation ($\rho_1 = 0.52$ vs. a strategy-peer median of 0.08), 25% to the use of a small, non-Big-Four auditor with two prior client sanctions, 20% to deteriorating readability in successive Form ADV brochures, and 20% to the fund's peripheral position in the co-investment network. This feature-level decomposition maps directly onto the investigative categories that SEC examiners are trained to evaluate, bridging the gap between algorithmic output and regulatory workflow.

LIME. Local Interpretable Model-agnostic Explanations (Ribeiro et al., 2016) construct a locally faithful interpretable model—typically a sparse linear model—in the neighborhood of a specific prediction. While SHAP provides exact feature attributions for tree-based models, LIME is model-agnostic and can be applied to any classifier, including deep neural networks and GNNs. For case-by-case investigation, LIME can generate explanations tailored to the specific characteristics of the flagged fund, highlighting the features that are most influential in its local decision region.

Attention visualization. For transformer-based and graph attention network models, the learned attention weights provide an intrinsic form of explainability. In a temporal transformer applied to monthly return series, attention weights reveal which historical time periods the model considers most relevant to the current fraud assessment—for example, concentrated attention on a six-month window during which returns became anomalously smooth. In a GAT applied to the fund-entity graph, attention weights reveal which relational connections—a specific auditor, a particular co-investor, a prior fund managed by the same principal—contributed most to the fraud score.

A worked example. Consider a hypothetical compliance alert generated by a deployed detection system. The alert identifies Fund XYZ, a mid-sized equity long/short fund, with a fraud probability of 0.78. The accompanying SHAP explanation reveals three primary contributors: (1) the fund's first-order serial correlation of $\rho_1 = 0.47$ is in the 98th percentile for equity long/short funds, suggesting return smoothing inconsistent with the fund's liquid equity mandate; (2) the fund recently switched from a Big-Four auditor to a regional firm whose prior clients include two funds that subsequently faced SEC enforcement actions; and (3) natural language analysis of the fund's most recent Form ADV brochure reveals a 40% increase in the Fog readability index relative to the prior filing, with increased use of hedging language ("may," "could," "potentially") in the risk factor section. An investigator reviewing this alert can immediately identify three independent lines of inquiry—return analysis, operational due diligence, and filing review—each supported by specific, quantified anomalies. This structured explanation transforms an opaque probability score into an actionable investigation plan.

3.6 Stage 5: Deployment and Monitoring

The final pipeline stage addresses the operational realities of running a fraud detection system in production. The transition from a research prototype that performs well on historical data to a deployed system that delivers reliable surveillance over time introduces challenges that the preceding stages do not address: processing cadence, model degradation, alert management, and integration with human investigative workflows.

Real-time versus batch processing. Hedge fund fraud detection operates on a fundamentally different temporal cadence than most other financial surveillance applications. Banking fraud detection systems process transactions in real time, rendering verdicts in milliseconds. Credit card fraud systems evaluate each swipe at the point of sale. Hedge fund surveillance, by contrast, is inherently batch-oriented: the primary input data—monthly returns and quarterly filings—arrive on fixed schedules, and the detection objective is to identify suspicious funds for further investigation rather than to block individual

transactions. The processing architecture should therefore be optimized for monthly or quarterly batch runs that score the full fund universe, supplemented by event-triggered evaluations when alternative data signals (e.g., breaking news, sudden web traffic changes, litigation filings) indicate the need for inter-batch reassessment. This hybrid batch-plus-event architecture matches the temporal structure of the available data while reducing the detection lag discussed in Section 2.2.3.

Concept drift detection. Models trained on historical data degrade over time as the distribution of fund behavior shifts—both because legitimate strategies evolve in response to market dynamics and because fraudsters adapt their methods to evade detection. Concept drift detection algorithms monitor the statistical properties of model inputs and outputs to identify when degradation occurs. The Adaptive Windowing (ADWIN) method maintains a variable-length window of recent observations and detects drift by comparing the distributions of the window's subsets. The Drift Detection Method (DDM) monitors the model's error rate and triggers an alarm when the error rate increases significantly relative to its historical baseline (Pang et al., 2021). Both methods can be adapted to the hedge fund context by monitoring fund-level anomaly scores, feature distributions, and model confidence scores across successive batch runs.

Retraining cadence and strategy. Drift detection triggers raise the question of when and how to retrain. Calendar-based retraining on a fixed schedule (e.g., semi-annually) provides predictability but may lag behind rapid distributional shifts. Drift-triggered retraining, which initiates model updates only when detected drift exceeds a threshold, is more adaptive but introduces the risk of frequent, potentially disruptive retraining cycles. A pragmatic hybrid approach combines scheduled retraining with drift-triggered emergency updates, maintaining a model registry that tracks the performance characteristics of each model version and enables rollback if a retrained model underperforms its predecessor.

Alert prioritization and workload management. A detection system that generates more alerts than investigators can process defeats its own purpose. Alert prioritization ranks flagged funds by a composite score that incorporates the model's fraud probability, the estimated financial exposure (assets under management as a proxy for potential investor losses), the novelty of the alert (is this a newly flagged fund or a persistent signal?), and the investigative tractability (are the data needed for follow-up investigation available?). This prioritization function transforms a raw list of flagged funds into a manageable investigation queue, enabling resource-constrained regulatory and compliance teams to allocate their attention to the cases with the highest expected value of investigation.

Human-in-the-loop integration. The deployment stage must be designed for human collaboration, not human replacement. Investigators who review alerts generate valuable feedback: confirmed fraud cases validate the model's predictions, cleared false positives identify systematic biases, and inconclusive cases highlight areas where the model's feature space is insufficient. This feedback loop, channeled back to the feature engineering and model training stages, enables continuous improvement. Active learning frameworks, in which the model selectively queries investigators about the cases most likely to improve its decision boundary, can maximize the information gained from each investigation and reduce alert fatigue—the well-documented tendency of human operators to discount or ignore alerts after experiencing repeated false positives. The EU AI Act's Article 14 human oversight requirement is directly satisfied by this design: the system informs human decision-making without replacing it, and investigators retain the ability to override model recommendations at every stage.

Integration with compliance infrastructure. In practice, a detection system does not operate in isolation. It must integrate with governance, risk, and compliance (GRC) platforms that manage regulatory reporting obligations, examination preparation, and enforcement coordination. API-based interfaces that deliver prioritized alerts, supporting explanations, and investigation packages to existing case management systems reduce

adoption barriers and ensure that detection outputs enter established workflows rather than creating parallel, disconnected processes. Audit trail functionality—logging every model inference, alert generation, investigator action, and feedback event—is essential for both regulatory compliance and system improvement.

This section has presented the five-stage detection pipeline that constitutes the paper's primary organizational contribution. The framework maps the fraud types catalogued in Section 2.1 to specific data sources, feature families, and AI methods, while identifying the explainability and deployment requirements that operational systems must satisfy. Section 4 builds on this framework by reviewing the specific studies that populate each stage, critically evaluating their methods, datasets, and findings within the context of hedge fund fraud detection.

4 Review of AI-Based Detection Methods

This section reviews the literature on AI and machine learning methods applied to hedge fund fraud detection and adjacent financial fraud domains. We organize the review around method families rather than chronologically, enabling direct comparison of approaches and systematic identification of gaps. For each family, we discuss the key studies, reported effectiveness, and limitations specific to the hedge fund context. Where direct evidence from hedge fund applications is unavailable, we draw on results from related financial fraud domains—accounting fraud, credit card fraud, and anti-money laundering—while carefully noting the assumptions required for transferability. The fraud taxonomy and data ecosystem established in Section 2 provide the evaluative lens: a method is assessed not only on its reported accuracy but on its applicability to the specific fraud types, data modalities, and operational constraints that characterize the hedge fund industry.

4.1 Classical Statistical and Rule-Based Approaches

The earliest detection methods for hedge fund fraud are rooted in forensic statistics and rule-based auditing procedures. These approaches predate the machine learning era but

remain foundational both as standalone screening tools and as feature generators for more complex models.

Benford's law, which predicts that the leading digit d in naturally occurring numerical data follows the distribution $P(d) = \log_{10}(1 + 1/d)$ (Benford, 1938), has been extensively applied to financial data forensics. Nigrini (2012) systematized its use for fraud detection, demonstrating that fabricated numerical data—because humans are poor intuitive generators of logarithmic digit distributions—tend to exhibit statistically significant deviations from Benford's expected frequencies. Applied retrospectively to Madoff's reported returns, automated digit-frequency tests flagged anomalies in nine out of ten statistical tests, lending credibility to the approach as a screening tool. However, Benford's law requires sufficiently large samples to achieve adequate statistical power, a condition that is rarely satisfied by individual hedge fund return series comprising at most 120–240 monthly observations. Moreover, a knowledgeable fraudster—aware that regulators employ Benford tests—can engineer return series that satisfy digit-frequency constraints while remaining fabricated in substance.

Serial correlation analysis represents a second classical pillar. Getmansky et al. (2004) developed a moving average model, MA(k), for hedge fund returns that decomposes observed returns into a true economic component and a smoothing component attributable to the managed pricing of illiquid assets. Their analysis demonstrated that 30–40% of hedge funds in the Lipper TASS database exhibit statistically significant positive serial correlation at lag one, consistent with return smoothing. While serial correlation alone does not imply fraud—illiquidity in legitimate portfolios produces similar signatures—the magnitude and persistence of serial correlation, particularly when inconsistent with a fund's stated strategy and asset class, constitutes a valuable fraud indicator. The contribution of this work to subsequent ML-based approaches is primarily methodological: the smoothing parameters estimated under the MA(k) model are now standard features in supervised classification systems.

Bollen and Pool (2012) advanced distributional analysis by identifying a "kink" at zero in the return distributions of a subset of hedge funds—a discontinuity in which

small positive returns appear with significantly higher frequency than small negative returns. This pattern, interpreted as evidence that managers selectively defer or reclassify marginal losses, correctly identified approximately 50% of funds that subsequently faced SEC enforcement actions. Brown et al. (2008) developed an operational risk scoring approach, the omega-score, derived from Form ADV data that captures governance and organizational risk factors; funds scoring above the threshold experienced significantly higher failure and regulatory action rates. Dimmock and Gerken (2012) applied logistic regression to SEC filing data, demonstrating that past regulatory violations, ownership structures, and custody arrangements predict future fraud with an AUC in the range of 0.65–0.70.

The collective limitation of these classical approaches is their specialization: each detects a specific statistical signature of a specific fraud type. Benford tests identify digit manipulation; serial correlation analysis detects return smoothing; distributional discontinuity analysis flags selective loss avoidance. None captures the multi-dimensional, heterogeneous patterns that characterize sophisticated fraud, and all exhibit high false positive rates when deployed independently. Their enduring value lies in their complementarity: the features they generate—serial correlation coefficients, Benford test statistics, distributional shape parameters, operational risk scores—form the bedrock of the feature engineering stage in modern ML pipelines (see Section 3).

4.2 Tree-Based and Ensemble Methods

Tree-based ensemble methods currently dominate tabular fraud detection across the financial industry, and their strengths align particularly well with the hedge fund context. Random forests (Breiman, 2001) aggregate predictions from hundreds of independently grown decision trees, each trained on a bootstrap sample with random feature subsets, yielding models that are robust to overfitting, tolerant of missing data, and capable of handling mixed feature types—numerical return statistics alongside categorical variables such as strategy classification and auditor identity—without requiring feature standardization.

The advent of gradient boosting machines, particularly XGBoost (Chen and Guestrin, 2016), LightGBM (Ke et al., 2017), and CatBoost (Prokhorenkova et al., 2018), elevated ensemble performance further. These algorithms construct trees sequentially, with each tree correcting the residuals of its predecessors, achieving state-of-the-art accuracy on tabular financial data across numerous benchmarks. Their built-in mechanisms for handling class imbalance—sample weighting, stratified subsampling, and cost-sensitive learning—are directly relevant to the skewed class distributions inherent in fraud detection.

The most rigorous large-scale application of ensemble methods to financial fraud detection is the work of Bao et al. (2020), who applied a RUSBoost ensemble—combining random undersampling with AdaBoost—to detect accounting fraud in publicly traded US firms. Training on over 28,000 firm-year observations linked to SEC Accounting and Auditing Enforcement Releases (AAERs), their model achieved an AUC of 0.725, substantially outperforming the logistic regression benchmark of Dimmock and Gerken (2012). Although this study targeted accounting fraud rather than hedge fund fraud specifically, the methodological template—ensemble learning on SEC enforcement labels with systematic feature engineering from regulatory filings—is directly transferable. More recent work on stacking ensembles that combine XGBoost, LightGBM, and CatBoost through a meta-learner has reported F_1 scores approaching 0.88 in financial fraud detection tasks, the highest among individual method families in the broader fraud detection literature (Hilal et al., 2022).

A critical strength of tree-based ensembles for the hedge fund domain is their compatibility with modern explainability methods. SHAP (SHapley Additive exPlanations) values, for which Lundberg and Lee (2017) provide exact computation algorithms for tree-based models, decompose each prediction into per-feature contributions, enabling explanations of the form: "this fund was flagged because its serial correlation at lag one is 2.3 standard deviations above the strategy-peer median, its Sharpe ratio is implausibly high given its stated volatility, and its auditor has been associated with two previously sanctioned funds." Such explanations satisfy the transparency requirements discussed in Section 5 and align with regulatory expectations for evidence-based enforcement.

The principal limitation of tree-based methods in this context is their reliance on supervised training with labeled fraud data. With only 50–100 confirmed hedge fund fraud cases across the historical record of SEC enforcement, the effective training set for the positive class is perilously small. Ensemble methods can mitigate class imbalance through resampling and cost-sensitive weighting, but they cannot overcome the fundamental heterogeneity of the positive class: the statistical signature of a Ponzi scheme differs qualitatively from that of NAV manipulation or style drift, and a model trained on pooled fraud labels may learn a compromise decision boundary that fails to detect any individual fraud type with high sensitivity.

4.3 Deep Learning Approaches

Deep learning architectures offer the potential to capture complex, nonlinear patterns in hedge fund data that tree-based methods may miss, but their application to this domain is constrained by data scarcity and opacity.

Long Short-Term Memory (LSTM) networks (Hochreiter and Schmidhuber, 1997) process sequential data through gated memory cells that selectively retain or discard information across time steps, making them naturally suited to the analysis of monthly return series. LSTMs can detect temporal patterns indicative of return smoothing—gradual shifts in serial correlation structure, time-varying volatility suppression, and regime-dependent anomalies—that fixed-window statistical tests cannot capture. The sequential nature of LSTM processing aligns with the temporal dynamics of hedge fund fraud, which typically escalates over months or years rather than occurring as a single discrete event.

Convolutional neural networks (CNNs), originally developed for image recognition (LeCun et al., 2015), have been adapted for financial time series analysis through the encoding of return sequences as two-dimensional representations—for example, mapping monthly returns to Gramian Angular Fields or recurrence plots—that CNN architectures can process using spatial pattern recognition. Hybrid CNN-LSTM architectures that apply convolutional layers for local feature extraction followed by recurrent layers for

temporal aggregation have shown promise for multi-scale temporal analysis, capturing both short-term return anomalies and longer-term behavioral shifts.

Transformer architectures (Vaswani et al., 2017), which employ self-attention mechanisms to model dependencies across arbitrary positions in a sequence, have emerged as a powerful alternative to recurrent models for sequential data. In the hedge fund context, transformers offer a specific advantage: their attention mechanism can capture long-range dependencies across sparse monthly return series—linking, for example, suspicious return patterns in a fund's first year of operation to anomalies that emerge three years later—without the vanishing gradient problems that limit LSTM effectiveness on long sequences.

Autoencoders represent the most directly applicable deep learning paradigm for hedge fund fraud detection under conditions of label scarcity. Trained to reconstruct input data through a compressed latent representation, autoencoders learn the distributional properties of normal fund behavior; funds with high reconstruction error—whose return characteristics deviate significantly from the learned normal distribution—are flagged as anomalous. Deep autoencoders have achieved an AUC of approximately 0.79 on hedge fund return data in anomaly detection frameworks (Chalapathy and Chawla, 2019), competitive with supervised approaches despite requiring no fraud labels. This unsupervised formulation sidesteps the labeled data bottleneck and avoids the data poisoning vulnerability that afflicts supervised methods (see Section 5.1).

The limitations of deep learning in this domain are substantial. Neural networks are data-hungry: modern architectures typically require thousands to millions of training examples to learn effective representations, yet the hedge fund universe comprises roughly 10,000–15,000 funds with at most 120–240 monthly return observations each. Overfitting is a pervasive risk, particularly for models with millions of parameters trained on datasets with fewer than 100 positive examples. Deep models are also opaque, producing predictions that resist human interpretation—a liability under the explainability requirements of the EU AI Act discussed in Section 5.3. Hyperparameter sensitivity compounds these concerns: small changes in architecture, learning rate, or regularization can produce

qualitatively different results, undermining reproducibility and the reliability of reported performance claims.

4.4 Natural Language Processing for Financial Filings

The text content of regulatory filings, offering memoranda, and investor communications constitutes a rich but underexploited data modality for hedge fund fraud detection. The evolution from bag-of-words representations to domain-specific transformer models has dramatically expanded the information that can be extracted from financial text.

Early NLP approaches to financial fraud relied on word-frequency features and domain-specific sentiment dictionaries. Loughran and McDonald (2011) demonstrated that general-purpose sentiment lexicons—such as the Harvard General Inquirer—perform poorly on financial text because many words with negative general sentiment (e.g., "liability," "tax," "depreciation") carry neutral or technical meaning in financial contexts. Their financial-domain sentiment dictionary, refined in subsequent work (Loughran and McDonald, 2016), provided more accurate sentiment classification for SEC filings and established the principle that financial NLP requires domain-adapted lexical resources.

The transformer revolution in NLP has produced models specifically adapted to financial language. FinBERT (Araci, 2019), a BERT model (Devlin et al., 2019) fine-tuned on a large corpus of financial news and communications, achieved 87% accuracy on financial sentiment classification tasks, substantially outperforming general-purpose models. SEC-BERT (Loukas et al., 2022), pre-trained specifically on SEC EDGAR filings, improved named entity recognition and document classification for regulatory texts by learning the distinctive vocabulary, syntactic patterns, and discourse structures of financial filings.

Applications to hedge fund fraud detection exploit several textual signals. First, vague or evasive language in strategy descriptions—excessive use of hedging phrases, abstract terminology, or circular definitions—may indicate an attempt to obscure the true nature of a fund's investment activities. Second, changes in filing complexity over time, measured through readability indices, sentence length distributions, and lexical diversity, have been associated with subsequent regulatory action: funds that progressively increase

the opacity of their disclosures may be concealing deteriorating performance or emerging fraud. Third, boilerplate deviation analysis identifies funds whose language departs from the standard templates used by legitimate funds; while modest deviation reflects genuine strategic differentiation, extreme deviation—or suspiciously precise conformity—may signal either carelessness or deliberate obfuscation.

Multi-modal fusion of NLP features with quantitative return data has been shown to improve detection over either modality alone, with reported improvements of approximately 3–5% in AUC (Ahmed et al., 2024). The logic is intuitive: a fund whose textual descriptions claim a conservative equity long/short strategy but whose return statistics exhibit factor exposures consistent with leveraged distressed credit presents a stronger fraud signal than either the textual or quantitative anomaly alone.

The limitations of NLP-based detection are both practical and structural. Regulatory filings are submitted quarterly or annually, introducing substantial detection lag. Filings are heavily boilerplate, with much of the text carrying minimal informational content—a low signal-to-noise ratio that challenges even sophisticated language models. Most critically, the standardized nature of financial filings means that fraudulent managers can—and do—match the expected linguistic patterns precisely, using compliance counsel and template providers to produce documents that pass textual screening while concealing substantive fraud. NLP is therefore most valuable as a complementary modality within a multi-modal detection framework rather than as a standalone screening tool.

4.5 Graph Neural Networks for Fund Networks

Graph neural networks (GNNs) represent a fundamentally different approach to fraud detection, operating on relational structures rather than individual entity features. In the hedge fund context, the relevant graph is the network of relationships among funds, fund managers, auditors, administrators, custodians, prime brokers, and investors—a heterogeneous, temporal graph whose topology carries fraud-relevant information that is invisible to methods operating on tabular data.

The foundational work of Kipf and Welling (2017) on graph convolutional networks

(GCNs) established that node classification in graphs can be performed by aggregating features from neighboring nodes through learned convolutional filters. Hamilton et al. (2017) introduced GraphSAGE, an inductive learning framework that generates embeddings for unseen nodes by sampling and aggregating features from a node's local neighborhood—a capability critical for scoring newly launched hedge funds that did not exist during model training (the cold-start problem discussed in Section 6.2.2). Graph attention networks (GATs), introduced by Veličković et al. (2018), apply attention mechanisms to weight the contributions of different neighbors, enabling the model to prioritize the most informative relationships.

The application of GNNs to financial fraud detection has yielded promising results. Wang et al. (2019) applied a semi-supervised GNN to transaction networks, achieving an AUC of 0.87 and demonstrating that graph-based features substantially outperform tabular features alone for network-embedded fraud. Liu et al. (2021) addressed the "camouflage" problem—in which fraudulent actors strategically surround themselves with legitimate connections to evade graph-based detection—using a GAT architecture with heterogeneous neighbor sampling that identifies fraud even when the immediate neighborhood appears benign.

The application of these methods to hedge fund networks is theoretically compelling but empirically underexplored. A hedge fund's service provider network—its auditor, administrator, custodian, and prime broker—constitutes a governance ecosystem whose composition correlates with fraud risk. Funds that employ small, unregistered auditors, lack independent administrators, or concentrate their service provider relationships within a small cluster of interconnected entities exhibit elevated fraud risk (Brown et al., 2008). Temporal knowledge graphs that track the evolution of these relationships over time can capture dynamic signals: a fund's sudden change of auditor, an administrator's simultaneous association with multiple subsequently sanctioned funds, or a manager's departure from one fund and rapid launch of another with the same service provider constellation.

The strengths of GNN-based approaches are distinctive: they capture relational and

structural information—guilt by association, network centrality, and structural equivalence—that is fundamentally inaccessible to methods operating on individual fund features. The primary limitations are practical. Graph construction for the hedge fund domain requires entity resolution across multiple databases—matching SEC filing entities to commercial database records to prime brokerage relationships—a data engineering challenge that is largely unaddressed in the literature. Relationship data are often incomplete: not all service provider relationships are disclosed in public filings, and the depth of relationships (e.g., the extent of an auditor's verification procedures) is unobservable. Computational cost scales with graph size and density, and training GNNs on the full hedge fund ecosystem graph, including temporal edges, demands substantial infrastructure.

4.6 Semi-Supervised and Self-Supervised Methods

The fundamental label scarcity problem in hedge fund fraud detection—fewer than 100 confirmed cases against a background of 10,000+ funds—motivates methods that can leverage the vast majority of unlabeled data. Semi-supervised and self-supervised approaches offer principled frameworks for doing so.

Label propagation and self-training algorithms extend sparse label information through the data manifold: a small number of labeled fraud and non-fraud examples propagate their labels to nearby unlabeled examples in feature space, iteratively expanding the labeled set. These methods are effective when fewer than 5% of the data carry labels, a condition that is characteristic of the hedge fund context. Self-training, in which a classifier iteratively labels its most confident predictions and retrains on the expanded set, can achieve substantial performance gains when the initial model is reasonably well-calibrated.

Contrastive learning represents a more recent and potentially more powerful paradigm. By learning representations that maximize agreement between differently augmented views of the same fund (positive pairs) while minimizing agreement between different funds (negative pairs), contrastive methods produce embeddings that separate normal from anomalous behavior without requiring explicit fraud labels (Pang et al., 2021). The

resulting representations can then be used for downstream classification with very few labeled examples, addressing both the label scarcity and the heterogeneity of the positive class.

Self-supervised pre-training on unlabeled fund returns—using objectives such as masked return prediction (predicting held-out months from context), temporal order prediction (determining whether a return sequence is in correct chronological order), or next-period forecasting—creates general-purpose representations of fund behavior that capture distributional regularities in the hedge fund universe. These representations can subsequently be fine-tuned for fraud detection with minimal labeled data, following the pre-train-then-fine-tune paradigm that has proven successful across NLP, computer vision, and other domains. Transfer learning from related financial domains—adapting models trained on banking fraud, insurance fraud, or accounting fraud to the hedge fund context—offers an additional avenue for overcoming data limitations, though the degree of transferability across fraud domains remains an empirical question.

The strength of these methods lies in their direct engagement with the label scarcity problem. Their limitation is sensitivity to distributional assumptions: performance degrades when the unlabeled data distribution shifts over time or when the labeled examples are not representative of the broader fraud population. Careful validation protocols—including temporal out-of-sample evaluation and robustness checks under distribution shift—are essential to guard against silent failure, in which a model appears to perform well on standard metrics while systematically missing emerging fraud patterns.

4.7 Synthetic Data and Data Augmentation

Data augmentation techniques address the class imbalance problem by generating synthetic examples of the minority class, expanding the effective training set for supervised and semi-supervised methods.

SMOTE (Chawla et al., 2002) remains the most widely used augmentation technique in financial fraud detection. By creating synthetic minority-class samples through interpolation between existing positive examples in feature space, SMOTE increases the

representation of the fraud class without duplicating existing examples. However, as discussed in the context of OP4 (Section 6.2.1), SMOTE's interpolation assumption is problematic for hedge fund fraud: interpolating between a Ponzi scheme and a valuation fraud generates synthetic examples that correspond to no plausible fraud pattern. Adaptive variants such as Borderline-SMOTE and ADASYN (Douzas et al., 2018), which concentrate synthetic sample generation near the decision boundary, partially address this issue but do not resolve the fundamental heterogeneity of the positive class.

Generative adversarial networks (GANs; Goodfellow et al., 2014) offer a more sophisticated approach to synthetic data generation. Conditional GANs, which condition the generation process on fraud type or other attributes, can produce synthetic fraud examples that preserve the distributional properties of specific fraud subtypes. Because GANs learn the data distribution implicitly through adversarial training, they can capture complex dependencies—between return statistics, filing characteristics, and operational features—that interpolation-based methods discard. Variational autoencoders (VAEs; Kingma and Welling, 2014) provide an alternative generative framework with better-calibrated uncertainty estimates, which is advantageous for fraud detection applications where the confidence of generated examples matters.

The practical application of these methods to the hedge fund domain involves augmenting the 50–100 known fraud cases to create training sets of viable size for supervised learning. However, synthetic data generation introduces a validation circularity: the generator learns to produce data that resembles the known fraud examples, but if the known examples are not representative of the full spectrum of hedge fund fraud, the synthetic data will perpetuate and amplify the biases of the original sample. Ensuring that synthetic fraud cases are realistic without assuming that we already know what fraud looks like is a fundamental methodological challenge.

Recent work has proposed synthetic benchmark datasets for fraud detection research as a means of addressing the lack of public data (Fiore et al., 2019). These benchmarks, generated through calibrated simulation rather than direct augmentation of proprietary data, can provide standardized evaluation resources while avoiding the confidentiality

constraints that prevent the release of real regulatory data. The development of hedge-fund-specific synthetic benchmarks, calibrated to the empirical properties of real hedge fund returns and fraud cases, remains an open priority (see OP1 in Section 6.1.1).

4.8 Critical Assessment of the Literature

The preceding subsections have documented a broad and growing body of work, but a candid assessment reveals significant structural weaknesses that limit the field's maturity and the reliability of its claims.

Reproducibility. The majority of studies in this domain use proprietary datasets—licensed commercial databases, internal regulatory records, or bespoke compilations—that are unavailable to other researchers. Results cannot be independently verified, and reported performance metrics must be taken on trust. This contrasts sharply with adjacent fields such as credit card fraud detection, where public benchmarks have enabled rigorous, reproducible comparison across hundreds of methods and research groups.

Benchmark gap. No standard benchmark dataset exists for hedge fund fraud detection. Each study assembles its own data, defines its own fraud labels, and reports results on non-overlapping fund populations. Cross-study comparison is therefore effectively impossible: a reported AUC of 0.79 in one study cannot be meaningfully compared to an AUC of 0.72 in another when the datasets, label definitions, feature sets, and evaluation protocols differ in every respect.

Domain specificity. The most impressive performance claims in the literature—F_1 scores above 0.85, AUC values above 0.90—originate from adjacent domains (credit card fraud, payment fraud, banking fraud) where labeled data are abundant and the statistical properties of fraud are well characterized. The transferability of these results to the hedge fund context, with its sparse data, heterogeneous fraud types, and sophisticated adversaries, is uncertain at best. Studies that report high performance on financial fraud benchmarks may overestimate effectiveness when applied to the specific conditions of

hedge fund surveillance (Bolton and Hand, 2002; Phua et al., 2010).

Class imbalance handling. The treatment of class imbalance varies enormously across studies and is often inadequately reported. Some studies apply SMOTE without evaluating its impact; others use cost-sensitive learning with ad hoc cost ratios; still others report only accuracy—a metric that is meaningless under severe class imbalance, as a model that classifies all funds as non-fraudulent achieves accuracy above 97%. The absence of standardized protocols for handling and reporting class imbalance makes performance claims across studies difficult to compare.

Evaluation protocols. The use of temporal train-test splits—training on data from period t and evaluating on data from period $t + 1$—is inconsistent across the literature. Many studies employ random cross-validation, which allows information from the future to leak into the training set and inflates reported performance. In a domain where fraud patterns evolve over time and concept drift is a known challenge, temporal evaluation is not a methodological refinement but a necessity; its absence in a substantial fraction of studies undermines confidence in reported results.

Publication bias. Positive results are preferentially published: studies reporting high detection rates are more likely to reach peer-reviewed venues than studies reporting null or modest results. The true performance landscape of AI-based hedge fund fraud detection is therefore likely less optimistic than the published literature suggests. Registered reports and pre-registered analysis plans, which commit to publication regardless of outcome, could partially address this bias but are not yet standard practice in this field.

Overfitting risk. With only approximately 50–100 labeled fraud cases in the historical record, even moderate-complexity models risk overfitting to the idiosyncratic characteristics of specific fraud schemes rather than learning generalizable fraud patterns. A model that achieves high performance on a held-out sample of, say, 20 fraud cases may simply have memorized the statistical fingerprints of the specific Ponzi schemes, valuation frauds, and style misrepresentations in its training set, without acquiring the capacity to detect

novel fraud types. This overfitting risk is amplified by the common practice of reporting results on a single train-test split rather than multiple independent evaluations.

Taken together, these concerns suggest that the field of AI-based hedge fund fraud detection is at an early stage of scientific maturity. Substantial progress in detection capability is likely achievable, but realizing this potential requires addressing the data, evaluation, and reproducibility challenges identified above. The research agenda presented in Section 6 proposes concrete steps toward this goal, beginning with the creation of standardized benchmarks (OP1) and the development of methods tailored to the extreme small-sample, heterogeneous-class conditions that define this domain (OP4).

5 Adversarial Robustness, Regulatory Readiness, and Ethical Considerations

The preceding sections have surveyed AI methods for hedge fund fraud detection on the assumption that the data-generating environment is stationary and non-hostile. In practice, neither assumption holds. Hedge fund managers who engage in fraud are not passive subjects of classification; they are sophisticated, quantitatively trained adversaries who actively adapt their behavior to evade detection. Simultaneously, regulators in multiple jurisdictions are imposing new requirements on AI systems used in high-stakes decision-making, creating legal constraints that detection models must satisfy. This section addresses three interconnected challenges that any operationally viable fraud detection system must confront: adversarial robustness against strategic manipulation (Section 5.1), regulatory explainability and compliance requirements (Section 5.3), and the ethical considerations that attend algorithmic surveillance of financial actors (Section 5.5).

5.1 Adversarial Threat Model

The adversarial machine learning literature, originating with the seminal work of Goodfellow et al. (2015) on adversarial examples in image classification and systematized by

Biggio and Roli (2018), has established that ML models are vulnerable to carefully crafted inputs designed to induce misclassification. In the financial fraud context, this vulnerability takes on a distinctive character. The adversary is not a script kiddie probing an image classifier with pixel-level perturbations; rather, the adversary is typically a PhD-level quantitative analyst with deep knowledge of statistical methods, access to the same academic literature that informs detection systems, and strong financial incentives to remain undetected. This qualitative difference in adversary capability fundamentally reshapes the threat model.

We identify four principal attack vectors relevant to hedge fund fraud detection.

Data poisoning. A fraudulent manager who reports fabricated returns to commercial databases effectively poisons the training data on which supervised detection models rely. Because hedge fund reporting to databases such as HFR and Lipper TASS is voluntary and largely unverified, a manager can engineer reported return series that appear statistically benign—satisfying Benford's law, exhibiting appropriate levels of serial correlation, and maintaining plausible distributional properties—while concealing the underlying fraud. **?** demonstrate that data poisoning attacks can degrade model performance by 5–12% even when the fraction of poisoned samples is small, and this finding is particularly concerning in the hedge fund context, where the base rate of detected fraud in training datasets is already low and class imbalance amplifies the impact of corrupted labels.

Evasion attacks. The most natural form of adversarial attack in this domain is evasion: structuring reported returns to avoid triggering detection thresholds. Return smoothing, which Getmansky et al. (2004) identified as a widespread practice among hedge funds investing in illiquid assets, is itself a form of adversarial noise injection—it transforms the observable return series to suppress volatility signals and serial correlation patterns that detection models rely upon. More sophisticated evasion strategies might involve optimizing reported returns against a known or estimated detection boundary, exploiting the fact that minor perturbations bounded by financial plausibility constraints can substan-

tially alter model predictions. Cartella et al. (2021) demonstrate that adversarial attacks using the Fast Gradient Sign Method (FGSM) of Goodfellow et al. (2015) and Projected Gradient Descent (PGD) of Madry et al. (2018) degrade AUC by 8–15% across financial fraud detection models. More broadly, recent work on adversarial robustness in financial ML reports a mean AUC degradation of 10.6% across surveyed detection systems, with even minor plausibility-bounded perturbations elevating calibration error and increasing expected portfolio loss by approximately 5%.

Model extraction. A sophisticated adversary can attempt to reverse-engineer a regulator's detection model by observing enforcement patterns over time. If a manager can infer which statistical features or behavioral patterns trigger regulatory scrutiny—by analyzing which funds are investigated, which enforcement actions are brought, and which anomalies are flagged in examination letters—the manager can construct a surrogate model of the detection system and optimize reported behavior to remain below its decision boundary. This model extraction attack is facilitated by the public nature of SEC enforcement actions and examination priority announcements, which inadvertently reveal information about the features and thresholds that regulators prioritize.

Strategic timing and regime exploitation. Hedge fund fraudsters can exploit temporal dynamics that most detection models do not account for. By timing fraudulent activity to coincide with periods of market stress—when legitimate fund returns exhibit unusual distributional properties—a fraudster can mask fabricated returns within the broader noise of market dislocation. Similarly, a manager can introduce fraudulent reporting gradually, allowing detection models trained on historical data to treat the evolving fraud signature as a legitimate regime change rather than an anomaly.

The practical significance of these threats is difficult to overstate. A survey of financial institutions found that 78% lacked formal adversarial resilience policies for their ML-based detection systems, suggesting that the gap between the adversarial threat landscape and institutional preparedness remains wide.

5.2 Defense Mechanisms

The adversarial ML literature offers several defense strategies, though their application to financial fraud detection requires careful adaptation.

Adversarial training. The most direct defense involves augmenting training data with adversarial examples, forcing the model to learn decision boundaries that are robust to perturbations. Madry et al. (2018) formalize this as a min-max optimization problem, where the inner maximization generates worst-case perturbations and the outer minimization trains the model to withstand them. **?** show that robust optimization applied to financial fraud detection models can recover 60–70% of the AUC lost to adversarial attacks, reducing attack success rates from approximately 35% to 5% in controlled experiments. However, adversarial training is computationally expensive—typically requiring 5–10 times the training cost of standard optimization—and raises the question of which perturbation model to use: ℓ_∞-bounded perturbations, which dominate the computer vision literature, may not capture the structured, financially constrained perturbations that a hedge fund fraudster would employ.

Ensemble diversity as implicit defense. Ensemble methods, which aggregate predictions from multiple heterogeneous base learners, provide a natural form of adversarial robustness. An adversary who optimizes an evasion strategy against one model component is unlikely to simultaneously fool all components, particularly when the ensemble incorporates diverse model families (e.g., tree-based models, neural networks, and statistical anomaly detectors) that define different decision boundaries in feature space. **?** demonstrate that a diversity metric computed over ensemble components correlates positively with robustness to adversarial perturbation, suggesting that the common practice of combining gradient-boosted trees with neural network classifiers may yield robustness benefits beyond the accuracy gains typically reported. This finding aligns with the broader ensemble learning literature, which has long recognized that diversity among base learners is a prerequisite for ensemble effectiveness.

Input validation and meta-level anomaly detection. A complementary defense strategy operates at the input level: rather than training the classification model itself to be robust, one can deploy a separate anomaly detection system that screens model inputs for signs of adversarial manipulation. In the hedge fund context, this meta-detection layer might flag return series that exhibit unusual statistical properties—such as suspiciously precise conformity to Benford's law, implausibly low serial correlation for the stated strategy type, or distributional characteristics that are inconsistent with the fund's reported asset class exposures. This layered approach treats adversarial detection as a distinct task from fraud detection, enabling specialized models for each.

Certified and randomized defenses. Certified defense methods, such as randomized smoothing, provide provable robustness guarantees within a specified perturbation radius. While theoretically appealing, these methods face practical limitations in the financial domain: the perturbation model must be defined over financially meaningful dimensions (returns, risk metrics, factor exposures) rather than abstract feature spaces, and the certification radius must be calibrated against plausible adversarial strategies rather than arbitrary ℓ_p norms. To date, no published work has adapted certified defense methods specifically to hedge fund fraud detection, representing an open area for future research.

5.3 Regulatory Explainability Requirements

The deployment of AI-based fraud detection systems operates within an evolving regulatory landscape that imposes substantive requirements on model transparency, interpretability, and human oversight. Two regulatory frameworks are particularly relevant: the European Union Artificial Intelligence Act and the guidance emanating from the U.S. Securities and Exchange Commission.

5.3.1 The EU Artificial Intelligence Act

The EU AI Act (Regulation 2024/1689), which entered into force in August 2024 with phased implementation through 2027, establishes the world's first comprehensive legal

framework for artificial intelligence (European Parliament and Council of the European Union, 2024). Financial fraud detection systems fall squarely within the Act's scope. Under Annex III, AI systems used for the assessment of creditworthiness and the evaluation of financial risk are classified as *high-risk AI systems*, subjecting them to the Act's most stringent requirements. Although the Act's text references creditworthiness and financial scoring most explicitly, the European Commission's interpretive guidance and the broad language of Articles 6 and 7 indicate that AI systems used to detect financial crime—including fraud detection models deployed by regulators, compliance departments, and financial institutions—will be subject to high-risk classification.

Three categories of requirements are particularly consequential for fraud detection systems. Article 13 mandates transparency: high-risk AI systems must be designed and developed in such a manner that their operation is sufficiently transparent to enable users to interpret the system's output and use it appropriately. For a fraud detection model, this implies that regulators or compliance officers who receive a fraud alert must be able to understand why the model flagged a particular fund—a requirement that favors models with inherent interpretability or robust post-hoc explanation capabilities. Article 14 requires human oversight: high-risk systems must be designed to allow effective oversight by natural persons during the period in which the system is in use, including the ability to decide not to use the system, to override its output, or to intervene in its operation. This requirement codifies the human-in-the-loop principle that many practitioners already advocate but that fully automated detection pipelines may violate. Article 9 mandates a comprehensive risk management system that identifies and evaluates foreseeable risks, including adversarial vulnerabilities, and establishes appropriate mitigation measures. The implication is that any AI-based fraud detection system deployed within the EU must not only perform accurately but must also demonstrate documented robustness against adversarial manipulation—linking the adversarial considerations of Section 5.1 directly to legal compliance.

5.3.2 SEC Regulatory Expectations

The United States has not enacted comprehensive AI legislation comparable to the EU AI Act. However, the SEC has signaled increasing regulatory attention to the use of AI and ML in investment management through multiple channels. The Division of Examinations has identified AI and emerging technology risk as a priority area in its annual examination priorities, and staff guidance has emphasized the fiduciary obligations of investment advisers who use algorithmic tools in portfolio management and risk assessment (**?**). The Division of Economic and Risk Analysis (DERA) itself employs ML models internally for market surveillance and enforcement targeting, creating implicit benchmarks for industry practice: if the SEC uses ML to detect fraud, it will develop institutional expectations about how such models should perform and what governance they require.

The SEC's recent enforcement actions against "AI washing"—bringing charges against firms that exaggerated or fabricated their use of AI in investment processes—signal that the Commission is attentive to the intersection of AI claims and investor protection. While these actions target firms that misrepresent AI capabilities rather than firms that deploy AI for detection, they establish that the SEC views AI governance as within its enforcement purview. Industry participants should anticipate that the SEC will eventually articulate more specific expectations for AI-based compliance and surveillance systems, particularly as the technology becomes more widespread in the investment management industry.

5.3.3 The Explainability–Performance Trade-off

The regulatory requirements described above create a fundamental tension with the technical performance characteristics of state-of-the-art detection models. The most accurate fraud detection approaches identified in the literature—deep ensemble methods, transformer-based architectures, and graph neural networks—are also the most opaque. These models learn complex, nonlinear decision boundaries across high-dimensional feature spaces, and their internal representations resist direct human interpretation. Conversely, inherently interpretable models—logistic regression, decision trees, and rule-based

systems—offer transparency that satisfies regulatory requirements but typically achieve lower detection rates and higher false negative rates (Rudin, 2019).

Post-hoc explainability methods offer a partial bridge between these poles. SHAP (SHapley Additive exPlanations) values, introduced by Lundberg and Lee (2017), provide theoretically grounded feature attribution scores that decompose a model's prediction into the contribution of each input feature. LIME (Local Interpretable Model-agnostic Explanations), proposed by Ribeiro et al. (2016), constructs local surrogate models that approximate the behavior of a complex model in the neighborhood of a specific prediction. Both methods have been widely adopted in financial ML applications and can generate explanations of the form "this fund was flagged because of unusually low serial correlation, high Sharpe ratio relative to peer funds, and irregular patterns in monthly return distributions." However, post-hoc explanations carry important limitations: they add computational cost, may not faithfully represent the model's true decision process (Rudin, 2019), and can be unstable across similar inputs (Arrieta et al., 2020). Guidotti et al. (2019) provide a comprehensive taxonomy of explainability methods and emphasize that no single approach satisfies all desiderata—fidelity, stability, comprehensibility, and computational efficiency—simultaneously.

No regulatory consensus currently exists on what constitutes "sufficient" explainability for AI-based financial fraud detection. The EU AI Act's transparency requirements are formulated at a high level of abstraction, and their operationalization will depend on implementing acts, harmonized standards, and supervisory practice that have not yet fully materialized. This regulatory uncertainty itself constitutes a risk for organizations deploying detection systems, as compliance requirements may tighten retroactively. The development of domain-specific explainability benchmarks—tailored to the financial fraud detection context and aligned with regulatory expectations—represents an important open problem, which we address further in Section 6.

5.4 Readiness Assessment of Detection Method Families

The interplay among adversarial robustness, explainability, and regulatory compliance creates a multi-dimensional evaluation space in which no single method family dominates across all criteria. We assess the principal method families qualitatively along five dimensions: adversarial robustness, intrinsic explainability, post-hoc explainability quality, regulatory compliance readiness, and deployment maturity.

Linear and logistic models. Linear models occupy an extreme position in this space: they offer high intrinsic explainability, as coefficients map directly to feature importance, and they satisfy regulatory transparency requirements with minimal additional infrastructure. However, their adversarial robustness is low. Linear decision boundaries are trivially invertible—an adversary who knows or estimates the model coefficients can compute the minimal perturbation needed to cross the classification threshold. Their detection accuracy for complex, nonlinear fraud patterns is also limited, restricting their utility as standalone detection systems. They remain valuable as baseline models and as components within ensemble architectures.

Tree-based ensemble methods. Gradient-boosted decision trees (e.g., XGBoost, LightGBM) and random forests currently occupy the most favorable position across the readiness dimensions. Their ensemble structure provides moderate adversarial robustness, as evasion attacks must simultaneously fool multiple decision paths. They offer moderate intrinsic explainability through feature importance scores and partial dependence plots, and they are compatible with high-quality post-hoc explanations via SHAP, for which Lundberg and Lee (2017) provide exact computation methods for tree-based models. In terms of deployment maturity, gradient boosting dominates production fraud detection systems across the financial industry, benefiting from mature software ecosystems, well-understood hyperparameter tuning procedures, and extensive operational experience. Their regulatory compliance readiness is correspondingly high: compliance teams can present feature importance rankings, partial dependence plots, and individual SHAP explanations to regulators with reasonable confidence that these satisfy current

transparency expectations.

Deep learning models. Neural network architectures—including feed-forward networks, recurrent networks (LSTMs), autoencoders, and graph neural networks—offer the highest potential detection accuracy for complex fraud patterns but rank lowest on intrinsic explainability and face the greatest regulatory compliance challenges. Their adversarial robustness without explicit defense mechanisms is also low: neural networks are highly sensitive to adversarial perturbations, and the transferability of adversarial examples across network architectures means that an adversary who constructs an attack against one neural model may succeed against others (Goodfellow et al., 2015). Post-hoc explainability methods can be applied to deep models, but the quality and faithfulness of explanations tend to be lower than for tree-based models, and the computational overhead is higher. Organizations deploying deep learning for fraud detection should anticipate significant investment in explainability infrastructure and adversarial hardening to satisfy regulatory requirements.

Hybrid and ensemble architectures. Architectures that combine multiple model families—for example, using tree-based models for the primary classification decision and neural networks for feature extraction from unstructured data—can achieve favorable trade-offs across multiple readiness dimensions. The tree-based component provides an interpretable decision surface, while the neural component captures complex patterns that trees alone might miss. Ensemble diversity provides implicit adversarial robustness, and the modular structure facilitates selective explanation: regulators can be shown the interpretable classification layer, while the feature extraction layer is documented through its output characteristics rather than its internal mechanics. This hybrid approach is increasingly common in production financial ML systems and represents a pragmatic path toward regulatory compliance.

Anomaly detection methods. Unsupervised anomaly detectors—isolation forests, one-class SVMs, and autoencoder-based methods—occupy a distinctive position. They

do not require labeled fraud examples, circumventing the data poisoning vulnerability that afflicts supervised methods. Their adversarial robustness to evasion attacks is moderate: an adversary must remain within the learned distribution of normal behavior, which constrains the fraud strategies available. However, their explainability is variable. Isolation forests can identify which features contribute most to an anomaly score, but autoencoder reconstruction errors are difficult to decompose into feature-level explanations. Regulatory compliance readiness is moderate, limited primarily by the challenge of explaining why a fund's behavior deviates from the learned normal pattern in terms that non-technical stakeholders can evaluate.

5.5 Ethics and Algorithmic Bias

The deployment of AI-based surveillance systems for hedge fund fraud detection raises ethical considerations that extend beyond technical performance and legal compliance. We identify six concerns that merit careful attention from researchers, regulators, and practitioners.

First, the *harm from false positives* in this domain is substantial and asymmetric. A hedge fund falsely flagged as fraudulent may suffer severe reputational damage, investor redemptions, counterparty relationship termination, and regulatory scrutiny—consequences that can destroy a legitimate business even if the fraud allegation is ultimately unfounded. Market-level effects are also possible: if a prominent fund is publicly associated with a fraud investigation triggered by an AI system, the resulting market disruption may harm investors in other funds and destabilize related asset markets. The cost of false positives in hedge fund fraud detection is therefore categorically different from, say, false positives in email spam filtering, and detection systems must be calibrated accordingly.

Second, *selection bias from historical enforcement data* poses a systematic fairness concern. Supervised models trained on past SEC enforcement actions learn to detect patterns associated with historically prosecuted fraud, but the population of prosecuted cases is not a random sample of all fraud. Regulators have historically concentrated en-

forcement resources on certain fund types, strategies, and geographies—partly reflecting resource allocation decisions, partly reflecting the visibility and accessibility of different fraud schemes. Models trained on these biased enforcement histories may perpetuate and amplify existing patterns of selective scrutiny, systematically under-detecting fraud in overlooked categories while over-detecting in historically targeted ones.

Third, *fairness across fund characteristics* deserves explicit evaluation. Do detection models disproportionately flag small funds with limited operational infrastructure, funds investing in emerging markets with inherently higher return volatility, or funds managed by individuals from underrepresented demographic groups? These questions have received almost no empirical attention in the hedge fund fraud detection literature, yet they carry significant implications for equal treatment under regulatory enforcement. The development of fairness metrics adapted to the financial fraud detection context—where protected attributes may include fund size, geography, and strategy type rather than the demographic categories that dominate the algorithmic fairness literature—represents an important open challenge (**?**).

Fourth, the *dual-use nature* of fraud detection research creates an inherent tension. The same AI techniques that enable regulators and compliance teams to identify suspicious behavior can be repurposed by fraudulent actors to test and refine their evasion strategies. A published detection model, complete with feature definitions and decision thresholds, provides a roadmap for adversarial optimization. This dual-use concern is not hypothetical: the quantitative sophistication of hedge fund managers means that published research is readily consumed and operationalized.

Fifth, the *transparency-versus-gaming paradox* complicates the push for model explainability. Regulatory requirements for transparency—disclosing how detection models work, which features they use, and what thresholds they employ—directly conflict with the operational need to keep detection methods confidential. Every piece of information disclosed about a detection system is information that an adversary can exploit. Balancing regulatory transparency with adversarial security requires nuanced governance frameworks that current regulations do not adequately address.

Sixth, we recommend that organizations deploying AI-based fraud detection systems adopt a governance framework that includes *regular bias audits* across fund characteristics and demographic dimensions, *diverse and representative training data* that corrects for historical enforcement biases where possible, *human-in-the-loop decision-making* in which AI systems inform but do not replace human judgment in enforcement decisions, and *transparent model governance* with documented procedures for model development, validation, monitoring, and retirement. These principles align with emerging best practices in responsible AI deployment (Arrieta et al., 2020; Molnar, 2020) and with the human oversight requirements of the EU AI Act discussed in Section 5.3.1.

6 Research Agenda and Open Problems

The preceding sections have documented a field that is simultaneously promising and immature. AI methods can detect statistical anomalies in hedge fund returns, extract fraud-relevant signals from regulatory filings, and model relational structures among market participants—yet the literature remains fragmented, the datasets are proprietary or nonexistent, and no deployed system has been rigorously evaluated against adversarial manipulation by sophisticated fund managers. This section crystallizes the most pressing gaps into ten concrete open problems, organized into three categories: data challenges (OP1–OP3), methodological challenges (OP4–OP7), and deployment challenges (OP8–OP10). For each problem, we articulate why it is uniquely difficult in the hedge fund context, suggest methodological approaches, and outline evaluation criteria. Collectively, these problems constitute Contribution C3 of this paper: an actionable research roadmap designed to guide both academic researchers and industry practitioners toward high-impact work.

6.1 Data Challenges

Progress in AI-based hedge fund fraud detection is fundamentally constrained by data availability. Unlike adjacent domains such as credit card fraud, where large-scale public

benchmarks have catalyzed rapid methodological advancement, the hedge fund domain lacks standardized datasets, suffers from jurisdictional fragmentation, and relies on reporting cycles that introduce detection lags measured in months rather than milliseconds. The three problems in this category address these structural data deficiencies.

6.1.1 OP1: Benchmark Dataset Creation

Problem. No public benchmark dataset exists for hedge fund fraud detection. The credit card fraud community benefits from widely used benchmark datasets that have enabled reproducible comparison of hundreds of methods over the past decade. The financial statement fraud literature can draw on publicly available accounting data linked to SEC enforcement actions (Dimmock and Gerken, 2012). Hedge fund fraud detection, by contrast, has no comparable resource: each study assembles its own proprietary dataset, uses different fraud labels, and reports results on non-overlapping fund populations, making cross-study comparison effectively impossible.

Why this is uniquely challenging for hedge funds. The obstacles to benchmark creation are structural, not merely logistical. Fund-level return data are proprietary, maintained by commercial database vendors under restrictive licensing agreements that prohibit redistribution. The data are sparse—monthly returns rather than the millions of daily transactions available in payment fraud—and the total population is small, comprising roughly 10,000 to 15,000 funds in the major databases at any given time. Confirmed fraud labels are scarce and ambiguous: SEC enforcement actions capture only detected fraud, introducing survivorship bias in the label space itself, and the boundary between aggressive but legal return management and fraudulent misrepresentation is often adjudicated only after years of litigation.

Suggested approach. We propose a two-track strategy. The first track involves constructing a synthetic benchmark that combines realistic return generation with injected fraud patterns. Regime-switching models calibrated to empirical hedge fund return distributions (Getmansky et al., 2004) can generate plausible non-fraudulent return series, while fraud patterns—Ponzi-like return fabrication, NAV smoothing, and style drift—

can be injected at rates and with characteristics calibrated to known enforcement cases. The second track involves developing an anonymization and differential privacy protocol that would enable regulators who possess enforcement-labeled datasets to release sanitized versions for research use. Generative models, including variational autoencoders (Kingma and Welling, 2014) and generative adversarial networks (Goodfellow et al., 2014), trained on real regulatory data could produce synthetic datasets that preserve aggregate statistical properties—return distributions, cross-sectional correlations, and temporal dynamics—while provably protecting the identities of individual funds. Evaluation should be conducted on held-out synthetic fraud cases using out-of-sample detection metrics, with the benchmark designed to support multiple fraud types, class imbalance levels, and detection horizons.

6.1.2 OP2: Cross-Jurisdictional Data Integration

Problem. Hedge fund data are fragmented across regulatory jurisdictions, and no unified data infrastructure connects the reporting regimes of different national authorities. A fund domiciled in the Cayman Islands, managed from New York, with a European marketing passport and Asian prime brokerage relationships generates regulatory filings across multiple jurisdictions, each with different formats, frequencies, and disclosure requirements. An AI system operating within any single jurisdiction sees only a partial picture of the fund's operations.

Why this is uniquely challenging for hedge funds. This fragmentation is qualitatively different from the cross-border challenges in banking or payment fraud. Banks operating internationally are subject to consolidated supervision under frameworks such as Basel III, and payment networks like SWIFT provide standardized messaging that enables cross-border transaction monitoring. Hedge funds, by contrast, exploit regulatory arbitrage deliberately: domiciliation in offshore centers such as the Cayman Islands, the British Virgin Islands, Luxembourg, or Ireland is chosen precisely because these jurisdictions impose lighter reporting requirements. The SEC's Form ADV captures US-registered advisers, the UK Financial Conduct Authority (FCA) maintains its own registry, and

the European Securities and Markets Authority (ESMA) collects data under AIFMD—but these datasets are not linked, use different entity identifiers, and apply different classification schemes for fund strategies and risk metrics.

Suggested approach. Federated learning frameworks offer a promising path forward, enabling multiple regulatory agencies to train a shared detection model without exchanging raw data. Each jurisdiction would train local model updates on its own regulatory data, and a central aggregation server would combine these updates to produce a global model that benefits from the combined information across all participating jurisdictions. Differential privacy guarantees can be layered onto the federated protocol to provide formal privacy bounds. The principal research challenges are twofold: first, developing entity resolution methods that link the same fund or manager across jurisdictions without sharing identifiable data; and second, handling the heterogeneity of reporting formats through standardized feature extraction pipelines that map jurisdiction-specific filings into a common representation space.

6.1.3 OP3: Real-Time Alternative Data Pipelines

Problem. Most existing detection approaches operate on monthly or quarterly data, introducing a detection lag that sophisticated fraudsters actively exploit. By the time a suspicious pattern becomes visible in a fund's monthly return series or quarterly regulatory filing, months of additional investor capital may have flowed into the fund, and the manager may have had ample opportunity to adjust behavior or abscond with assets.

Why this is uniquely challenging for hedge funds. The detection lag in hedge fund surveillance is orders of magnitude greater than in other fraud domains. Banking fraud detection operates on real-time transaction streams with sub-second latency. Credit card fraud systems evaluate each transaction at the point of sale. Hedge fund surveillance, by contrast, depends on returns that are reported monthly with a typical lag of 30 to 60 days, regulatory filings that are updated annually or semi-annually, and holdings data that are disclosed quarterly with a 45-day delay. This temporal gap is not an implementation limitation but a structural feature of the hedge fund reporting regime.

Suggested approach. The integration of real-time alternative data with periodic fund data offers a route to substantially reducing detection lag. News sentiment analysis using transformer-based language models (Devlin et al., 2019) can flag emerging concerns about specific funds or managers within hours of publication. Social media monitoring can detect investor complaints, whistleblower signals, and reputational deterioration in near-real time. Web scraping of fund marketing materials, employee turnover on professional networks, and litigation filings can provide early warning signals that precede changes in reported returns by months. The research challenge lies in developing fusion architectures that coherently integrate these high-frequency, noisy alternative data streams with the low-frequency, high-reliability periodic data that form the backbone of existing detection models. Attention-based architectures that learn to weight different data modalities according to their informativeness at different time horizons represent a promising direction (Vaswani et al., 2017).

6.2 Methodological Challenges

Even with adequate data, the hedge fund fraud detection problem poses methodological challenges that distinguish it from fraud detection in other financial domains. The extreme rarity and heterogeneity of fraud events, the cold-start problem for new funds, temporal non-stationarity, and the need to fuse radically different data modalities each require tailored solutions that go beyond standard machine learning practice.

6.2.1 OP4: Extreme Class Imbalance at Small Scale

Problem. Hedge fund fraud detection faces a class imbalance problem that is qualitatively different from—and more severe than—the imbalance encountered in most other fraud detection domains. Standard oversampling techniques, developed for settings with abundant transaction data, fail to address the fundamental statistical challenges of the hedge fund context.

Why this is uniquely challenging for hedge funds. Credit card fraud detection, the canonical class imbalance problem in the ML literature, operates at a base rate of ap-

proximately 0.1–0.5% but compensates with millions or tens of millions of transactions, yielding thousands of positive examples for training. Hedge fund fraud has a comparable or lower base rate—estimates range from 0.5% to 3% depending on the fraud definition and time window—but operates on a total population of roughly 10,000 to 15,000 funds, yielding perhaps 50 to 100 labeled fraud cases across the entire historical record of enforcement actions. This is a small-N, high-dimensional, multi-modal problem. Standard oversampling methods such as SMOTE (Chawla et al., 2002) fail not merely because of imbalance but because the minority class is fundamentally heterogeneous: Ponzi schemes, NAV manipulation, and style drift produce entirely different statistical signatures, and interpolating between a Madoff-type fabrication and a Platinum Partners-type valuation fraud generates synthetic examples that correspond to no real fraud pattern. The problem is not too few fraud cases in aggregate; it is too few fraud cases *per fraud type*, with types that are qualitatively distinct.

Suggested approaches. Several research directions merit investigation. Few-shot learning methods, which learn to classify from very small numbers of examples by leveraging meta-learned representations, may enable fraud-type-specific detection with as few as five to ten labeled cases per type. Meta-learning across fraud types—training on a distribution of fraud detection tasks rather than a single classification problem—could produce models that generalize to novel fraud patterns from minimal examples. Semi-supervised contrastive learning, which learns representations by contrasting normal and anomalous funds in an embedding space without requiring hard labels for all examples, offers a way to exploit the large number of unlabeled funds. Transfer learning from related fraud domains (insurance fraud, financial statement fraud, money laundering) may provide useful inductive biases, though the degree of transferability across fraud domains remains an empirical question. Evaluation protocols must account for this heterogeneity: standard cross-validation on a pooled fraud class is insufficient, and researchers should adopt fraud-type-stratified splits that assess whether a model trained on known fraud types can detect held-out types.

6.2.2 OP5: Cold-Start Detection for New and Emerging Funds

Problem. Newly launched hedge funds lack the historical return data that serve as the primary input to most detection models. A fund with three months of track record provides insufficient statistical power for distributional tests, serial correlation analysis, or style-based anomaly detection. Yet the early period of a fund's life is precisely when fraud risk may be highest, as managers seek to establish credibility and attract capital.

Why this is uniquely challenging for hedge funds. The cold-start problem in hedge fund detection differs fundamentally from cold-start problems in other domains. In banking, a new account inherits the institutional history and risk profile of the customer who opens it. In e-commerce fraud, a new user can be assessed through device fingerprinting, behavioral biometrics, and graph-based features from the first interaction. A new hedge fund, by contrast, is genuinely informationally opaque: it has no return history, no regulatory filing history, and no track record against which claims can be validated. Moreover, the hedge fund industry's incubation structures—in which managers operate "paper portfolios" or seed-stage vehicles before formally launching—allow selective reporting of only successful track records, a phenomenon closely related to backfill bias (Agarwal et al., 2011; Fung and Hsieh, 2009). The fund that appears in a database may be the survivor of multiple failed incubation attempts, and the reported inception-date returns may represent a cherry-picked history.

Suggested approaches. Detection for new funds must rely on non-performance features. Operational due diligence characteristics—auditor quality, administrator independence, custody arrangements, and governance structures—are available from Form ADV at the time of registration and have demonstrated predictive power for subsequent fraud (Brown et al., 2008; Dimmock and Gerken, 2012). Transfer learning from funds with similar strategy descriptions, manager backgrounds, and organizational structures could provide prior distributions for expected return behavior, against which early realized returns can be evaluated. Network analysis of the manager's prior fund relationships—previous launches, co-manager affiliations, service provider networks—may reveal risk patterns that are invisible in the fund's own data. NLP applied to the fund's launch documents,

offering memoranda, and marketing materials can assess the plausibility and consistency of stated strategies before any returns are generated. The research challenge is to develop models that integrate these heterogeneous non-performance signals into a coherent risk score that is calibrated against the base rate of fraud in new funds specifically.

6.2.3 OP6: Temporal Concept Drift and Adaptive Models

Problem. Fraud patterns evolve over time as perpetrators learn from detected schemes and adapt their methods to evade current detection approaches. A model trained on historical fraud cases may become progressively less effective as the distribution of fraud signals shifts, a phenomenon known as concept drift. In standard classification settings, concept drift is addressed through periodic retraining or online learning algorithms. In the hedge fund context, however, the problem is compounded by a source of legitimate drift that is absent in most other fraud domains.

Why this is uniquely challenging for hedge funds. Hedge funds legitimately change their investment strategies in response to market conditions. A fund that begins as an equity long/short vehicle may pivot toward global macro during a period of heightened currency volatility, or a quantitative fund may shift from momentum to mean-reversion factors as market regimes change. These legitimate strategy transitions produce statistical signatures—changes in factor exposures, return distributions, and risk profiles—that closely resemble the signals associated with fraudulent strategy misrepresentation. In credit card fraud, spending pattern drift is slow and predictable, driven by lifecycle events and inflation. In banking fraud, behavioral drift reflects changes in channel usage and product adoption. Hedge fund drift, by contrast, can be sudden, large, and strategically motivated, making it fundamentally harder to distinguish legitimate adaptation from fraudulent concealment.

Suggested approaches. Strategy-aware drift detection algorithms that incorporate information about market regimes and factor conditions represent a promising direction. Online learning methods with explicit drift detectors—such as the Adaptive Windowing (ADWIN) algorithm or the Drift Detection Method (DDM)—can be augmented with

regime-switching models that identify whether observed changes in a fund's behavior are consistent with concurrent market dynamics or represent anomalous deviation. Factor-conditioned anomaly scores that normalize a fund's return characteristics against the behavior of peer funds pursuing similar strategies could reduce the false positive rate generated by legitimate strategy evolution. The key evaluation challenge is the construction of test sets that include both genuinely drifting fraud and legitimately adapting non-fraud, requiring either carefully annotated historical data or synthetic scenarios that embed both types of change.

6.2.4 OP7: Multi-Modal Fusion Architectures

Problem. Hedge fund fraud detection requires integrating information from multiple data modalities with fundamentally different characteristics: numerical return series, unstructured text from regulatory filings, high-dimensional portfolio snapshots, and graph-structured relational data. Each modality has different sampling frequencies, dimensionalities, and missing-data patterns. Existing approaches typically operate on a single modality or, at best, concatenate hand-engineered features from multiple sources, leaving substantial information on the table.

Why this is uniquely challenging for hedge funds. No other financial entity generates the specific combination of data modalities that characterizes a hedge fund. Monthly returns are sparse, low-frequency time series with as few as 36 to 120 observations per fund. Quarterly regulatory filings are unstructured narrative documents that require NLP processing. Form 13F holdings are high-dimensional portfolio snapshots reported with a 45-day lag. Network data—prime broker relationships, auditor connections, manager affiliations—are inherently relational. Alternative data sources such as news sentiment and litigation records arrive at irregular, high-frequency intervals. Fusion must handle these radically different frequencies, dimensionalities, and missingness patterns simultaneously. This stands in contrast to, for example, e-commerce fraud, where all relevant signals (transaction amount, device, location, history) are available at the same granularity for every event.

Suggested approaches. Attention-based multi-modal transformer architectures (Vaswani et al., 2017) that learn to weight different modalities dynamically according to their informativeness offer a principled framework for fusion. Cross-modal contrastive learning, which trains representations such that consistent signals across modalities are embedded nearby while inconsistencies are flagged, could specifically target the detection of funds whose textual descriptions contradict their quantitative behavior. Hierarchical fusion with modality-specific encoders—for example, recurrent networks for return series, pre-trained language models (Devlin et al., 2019) for text, and graph neural networks (Hamilton et al., 2017; Kipf and Welling, 2017) for relational data—can capture the inductive biases appropriate to each data type before combining representations at a higher level. The evaluation protocol should assess not only aggregate detection performance but also the marginal contribution of each modality, enabling researchers and practitioners to understand which data sources are most valuable and where investment in data acquisition would yield the greatest detection improvement.

6.3 Deployment Challenges

Even a technically superior detection model is of limited practical value if it cannot withstand adversarial manipulation, explain its decisions to regulators and courts, or integrate effectively into human investigation workflows. The three problems in this category address the gap between academic proof-of-concept and operational deployment.

6.3.1 OP8: Adversarial Robustness Guarantees

Problem. Fraud detection is inherently an adversarial problem: the targets of detection actively seek to evade it. A detection model that achieves high accuracy on historical data may be rendered ineffective if a sophisticated manager reverse-engineers its decision boundary and adjusts reported behavior accordingly. Adversarial robustness—the ability of a model to maintain performance under deliberate input manipulation—is therefore a necessary condition for deployment, not merely a desirable property.

Why this is uniquely challenging for hedge funds. The adversarial dynamics in hedge

fund fraud detection are qualitatively different from those in other fraud domains. Credit card fraudsters, while numerous, are typically unsophisticated and operate at scale, relying on volume rather than precision. Hedge fund managers contemplating fraud are, by contrast, among the most quantitatively sophisticated actors in the financial system. They employ PhD-level quantitative analysts, have access to the same statistical and machine learning tools used to build detection models, and can afford to hire consultants who specialize in regulatory compliance and forensic accounting. The adversary in this domain is not a script kiddie but a well-resourced, highly educated professional who may be able to simulate detection models, identify their decision boundaries, and engineer return series that evade detection while maintaining the appearance of legitimacy. This makes the adversarial game qualitatively different and demands robustness guarantees that go beyond standard perturbation-based adversarial training.

Suggested approaches. Certified robustness bounds that provide formal guarantees on the maximum change in model output under bounded input perturbations offer a principled foundation. Game-theoretic adversarial modeling, in which the detection system and the fraudster are modeled as players in a repeated game with asymmetric information, can capture the strategic dynamics of hedge fund fraud more faithfully than standard adversarial training paradigms. Red-teaming exercises in which financial domain experts attempt to construct return series that evade specific detection models would provide empirical assessments of robustness. Robust ensemble methods that combine diverse model families—statistical tests, tree-based classifiers, deep networks, and network-based detectors—can increase the difficulty of evasion by requiring an adversary to simultaneously fool multiple independent detection mechanisms. Research in this area should evaluate robustness not only against generic adversarial perturbations but against economically meaningful manipulations: return series that satisfy standard statistical plausibility checks, filings that pass automated consistency validation, and network structures that mimic legitimate fund-of-funds arrangements.

6.3.2 OP9: Explainability Without Sacrificing Performance

Problem. Regulatory requirements for AI systems used in financial surveillance demand transparency: the EU AI Act requires that high-risk AI systems produce outputs that are "sufficiently transparent to enable deployers to interpret the system's output and use it appropriately," and SEC enforcement actions must be supported by evidence that can withstand legal challenge. Yet the most powerful detection methods—deep neural networks, ensemble models, graph neural networks—are precisely those that resist straightforward explanation. The field faces a tension between detection power and interpretive transparency that current post-hoc explainability methods do not fully resolve.

Why this is uniquely challenging for hedge funds. The explainability requirements for hedge fund fraud detection are more stringent than for most other AI applications in finance. In credit scoring, the explainability bar is well established by decades of regulatory guidance, adverse action notice requirements, and settled law. In hedge fund fraud detection, the regulatory framework is newer and less settled: there is no standardized examination protocol for AI-assisted surveillance, enforcement actions based on AI-generated evidence are largely untested in court, and the legal standards for admissibility of algorithmic suspicion vary across jurisdictions. Explainability must therefore satisfy multiple audiences simultaneously: SEC or FCA examiners who need to understand why a fund was flagged during a routine examination, enforcement attorneys who need to present evidence in administrative proceedings, and potentially judges and juries in contested civil or criminal cases. This is a materially higher bar than the compliance dashboards that suffice for standard anti-money laundering systems. Rudin (2019) argued forcefully that high-stakes domains should prefer inherently interpretable models over post-hoc explanations of black boxes; the hedge fund fraud context exemplifies this argument.

Suggested approaches. Inherently interpretable architectures such as neural additive models and explainable boosting machines offer a promising middle ground, providing nonlinear modeling capacity while maintaining per-feature transparency. Where complex models are necessary for performance, faithful distillation—training an interpretable

student model to approximate the decision boundary of a complex teacher model—can provide explanations that are provably consistent with the teacher's behavior. Post-hoc methods such as SHAP (Lundberg and Lee, 2017) and LIME (Ribeiro et al., 2016) provide local explanations for individual predictions, but their faithfulness to the underlying model has been questioned in adversarial settings (Arrieta et al., 2020). Attention-based explanations that highlight which return periods, filing passages, or network connections contributed most to a fraud score offer a natural explanation format for financial investigators. A critical research direction is the development of standardized explanation templates designed specifically for regulatory audiences—structured reports that map model outputs to the specific fraud indicators that examiners are trained to evaluate, bridging the gap between ML output and regulatory workflow.

6.3.3 OP10: Human-AI Collaboration in Fraud Investigation

Problem. No AI system will replace human judgment in hedge fund fraud investigation in the foreseeable future. Detection models generate alerts; human investigators must evaluate those alerts, conduct follow-up analysis, and make enforcement decisions. The effectiveness of the overall system therefore depends not only on model accuracy but on the quality of the human-AI interaction: how alerts are prioritized, communicated, and acted upon, and how investigator feedback is incorporated to improve future detection.

Why this is uniquely challenging for hedge funds. Hedge fund fraud investigation is resource-intensive and cognitively demanding. A single investigation may require months of document review, forensic accounting analysis, and expert consultation. False positives are not merely an inconvenience—they consume scarce examination resources that could be directed toward genuinely suspicious funds. At the same time, the consequences of false negatives are severe: undetected fraud can compound over years, as the Madoff case demonstrated. The base rate of fraud is low enough that even a model with apparently high precision may generate more false positives than true positives in absolute terms, a well-known consequence of applying high-accuracy classifiers to rare events. Investigator alert fatigue—the tendency to discount or ignore alerts after experiencing repeated false

positives—is a serious operational risk that can negate the benefits of even a technically capable detection system.

Suggested approaches. Active learning frameworks, in which the model selectively queries investigators about the cases most likely to improve its decision boundary, can simultaneously reduce alert volume and maximize the information gained from each investigation. Prioritized alert queues that present cases ranked by both fraud probability and estimated financial impact can help investigators allocate their time to the highest-value cases. Investigation-ready explanation packages—structured reports that include the model's fraud score, the contributing features with their SHAP values, relevant historical comparisons to known fraud cases, and suggested lines of inquiry—can reduce the time from alert to actionable investigation. Adaptive threshold management, which adjusts the alert generation threshold based on current investigator capacity and historical precision at different operating points, can maintain a sustainable alert volume. Feedback loops that allow investigators to label alerts as true positives, false positives, or inconclusive and retrain the model accordingly are essential for long-term system improvement. The research challenge is to evaluate these human-AI interaction designs in realistic operational settings, which may require collaboration with regulatory agencies willing to conduct controlled trials.

6.4 Prioritization and Path Forward

Not all open problems are equally urgent or equally tractable. We propose a prioritization based on two criteria: *impact*, defined as the degree to which solving the problem would unlock progress on other problems, and *feasibility*, defined as the degree to which current methods and data can support meaningful research. Table 3 summarizes this assessment.

Two problems stand out as critical preconditions. OP1 (benchmark dataset creation) is foundational because without a shared evaluation resource, the field cannot conduct reproducible comparisons, and progress on nearly every other problem—class imbalance methods (OP4), drift detection (OP6), fusion architectures (OP7), and adversarial robustness (OP8)—is impeded by the inability to benchmark against common data. OP4

Table 3: Prioritization of open problems by impact and feasibility. Problems marked as preconditions enable progress on dependent problems.

ID	Problem	Impact	Feasibility	Dependencies
OP1	Benchmark dataset	Critical	Medium	Precondition for OP4, OP6, OP7
OP4	Class imbalance at small N	Critical	Medium	Precondition for all supervised methods
OP9	Explainability	High	Medium–High	Required for OP10, regulatory deployment
OP5	Cold-start detection	High	Medium	Benefits from OP1
OP7	Multi-modal fusion	High	Medium	Benefits from OP1
OP8	Adversarial robustness	High	Low–Medium	Requires OP1 for evaluation
OP6	Concept drift	Medium–High	Medium	Benefits from OP1
OP10	Human-AI collaboration	Medium–High	Low	Requires regulatory partnerships
OP3	Real-time data pipelines	Medium	Low–Medium	Data access constraints
OP2	Cross-jurisdictional integration	Medium	Low	Requires multi-regulator coordination

(extreme class imbalance) is equally critical because the small-N, heterogeneous-positive-class nature of the problem invalidates the standard supervised learning assumptions that underpin most detection methods; any methodological advance that does not grapple with this reality will fail to translate from laboratory settings to operational deployment.

The feasibility assessment reveals an important structural constraint: the most impactful problems—OP2 (cross-jurisdictional integration) and OP10 (human-AI collaboration)—are also those that require institutional collaboration that no single research group can orchestrate. Regulatory agencies possess the enforcement-labeled data needed for benchmark creation (OP1), the cross-border relationships needed for federated learning (OP2), and the operational environments needed for human-AI evaluation (OP10). Academic researchers bring methodological expertise in deep learning, adversarial robustness, and explainability. The hedge fund industry contributes domain knowledge about strategy evolution, operational due diligence, and the practical realities of fund management.

Progress on this research agenda therefore demands a collaborative model that brings these three communities together—through shared data initiatives, regulatory sandboxes for AI testing, and joint research programs—because no single community possesses the data, methods, and operational context needed to address these problems in isolation.

7 Conclusion

The explosive growth of the hedge fund industry, combined with its structural opacity and light disclosure requirements, has created a regulatory environment in which fraud can persist undetected for years. The Madoff Ponzi scheme, which ran for at least two decades and resulted in $65 billion in stated losses, stands as a stark illustration of this reality. Traditional detection methods—human auditors, whistleblower reports, and univariate statistical tests—remain necessary but insufficient. The mismatch between regulatory capacity and industry scale, compounded by the cognitive limitations of human judgment and the sophistication of modern fraudsters, demands a fundamental expansion of the detection toolkit. Artificial intelligence and machine learning offer that expansion, providing scalable, multi-dimensional, and real-time fraud detection capabilities that cannot be achieved through manual analysis alone.

This survey has synthesized the scattered literature on AI-based hedge fund fraud detection into a coherent analytical framework, mapped existing methods to specific fraud types, evaluated their robustness and regulatory readiness, and articulated a concrete research agenda. We conclude by summarizing our three principal contributions and distilling key takeaways for practitioners, regulators, and researchers.

Summary of Contributions

Contribution C1: Detection Pipeline Taxonomy. The five-stage detection pipeline framework introduced in Section 3 provides a unified structure for researchers and practitioners to understand how raw data—return series, regulatory filings, alternative data, and relational networks—are transformed into actionable fraud assessments. By systematically

mapping fraud types to appropriate AI methods at each stage, the taxonomy addresses a critical gap in the existing literature, which has treated detection methods in isolation without reference to the end-to-end workflow required for operational deployment. The pipeline makes explicit that no single detection method covers all fraud types; instead, effective surveillance requires a multi-stage architecture with method selection guided by the specific fraud typology under investigation.

Contribution C2: Adversarial and Regulatory Readiness Assessment. Current AI methods face significant adversarial vulnerabilities and uncertain regulatory compliance. Our evaluation in Section 5 documented a mean AUC degradation of 10.6% under adversarial perturbations across reviewed methods, with some techniques exhibiting degradation exceeding 25% when targeted by informed adversaries. This finding reveals a fundamental mismatch: hedge fund managers contemplating fraud are among the most quantitatively sophisticated actors in the financial system, yet most detection models have not been tested against adversaries with this level of expertise. Simultaneously, emerging regulatory frameworks—particularly the EU Artificial Intelligence Act, which classifies fraud detection systems as high-risk and mandates transparency and explainability—impose constraints that many current methods do not satisfy. This dual assessment bridges the gap between the technical machine learning literature and the practical demands of regulators and compliance professionals, highlighting that detection performance on historical data is a necessary but not sufficient condition for deployment.

Contribution C3: Actionable Research Roadmap. The ten open problems articulated in Section 6 each address a gap that is uniquely challenging in the hedge fund context. The creation of benchmark datasets (OP1) and the resolution of extreme class imbalance at small scale (OP4) emerge as critical preconditions that unlock progress across nearly all other problems. Multi-modal fusion architectures (OP7), adversarial robustness guarantees (OP8), and explainability without sacrificing performance (OP9) represent methodological frontiers that require new technical approaches. Cross-jurisdictional data integration (OP2) and human-AI collaboration in investigation workflows (OP10) demand institutional partnerships among regulators, academic researchers, and industry

practitioners. Collectively, these problems constitute a research agenda designed to guide the field toward methods that are not only technically sophisticated but operationally deployable, legally compliant, and robustly resistant to strategic manipulation.

Key Takeaways for Practitioners

Hedge fund managers, fund-of-funds allocators, and institutional investors seeking to integrate AI-based fraud detection into their due diligence workflows should consider three principal findings.

First, ensemble methods offer the best current balance of performance, interpretability, and deployment readiness. Gradient boosting models, stacking ensembles, and random forests consistently achieve strong detection performance across diverse fraud types while maintaining substantially greater interpretability than deep neural networks. These methods can be augmented with post-hoc explainability techniques such as SHAP values to produce investigation-ready outputs that identify which specific features contributed to a fraud score. For organizations without extensive machine learning infrastructure, these techniques provide an accessible entry point that does not require specialized hardware or deep learning expertise.

Second, no single detection method covers all fraud types, and a multi-stage pipeline with method selection guided by fraud typology is essential. Performance fabrication and regulatory fraud are amenable to detection using return-based statistical tests and NLP on regulatory filings, respectively, with detection difficulty ratings of 3/5 and 2/5 (Section 2.1). Market manipulation and allocation fraud, by contrast, require order-level trade data and cross-account dispersion analysis that are typically unavailable to external investors, achieving difficulty ratings of 5/5 and 4/5. Practitioners must therefore tailor their detection architecture to the data they can access and the fraud types most relevant to their risk exposure. An operational due diligence system focused on Ponzi-like fabrication will emphasize serial correlation tests, Benford's law, and Sharpe ratio plausibility; a prime broker conducting allocation fraud surveillance will prioritize cross-account win-rate asymmetry and timestamp analysis.

Third, adversarial robustness testing should be standard practice given the sophistication of potential adversaries. Unlike credit card fraud, where adversaries are typically unsophisticated and operate at scale, hedge fund fraud may be perpetrated by actors with PhD-level quantitative expertise who can simulate detection models and engineer evasive strategies. The mean 10.6% AUC degradation documented in Section 5 understates the risk because the adversarial perturbations tested in the literature are generic, not tailored to the economic constraints of the hedge fund domain. Practitioners deploying detection systems should conduct domain-specific red-teaming exercises in which financial experts attempt to construct fraudulent return series that evade the deployed models, iteratively improving robustness before operational use.

Key Takeaways for Regulators

Financial regulators and supervisory authorities considering AI-based surveillance for hedge fund oversight should attend to two critical findings.

First, AI-based surveillance can address the scale mismatch between regulatory resources and industry size, but models must satisfy emerging explainability requirements. The SEC's Division of Examinations conducts only a fraction of possible inspections in any given year, creating long windows during which fraud can operate undetected. AI systems that process thousands of fund return series, regulatory filings, and alternative data sources simultaneously can flag suspicious patterns at a scale that human examiners cannot achieve. However, the EU AI Act's requirement that high-risk systems produce "sufficiently transparent" outputs poses a genuine constraint: deep neural networks and complex ensembles may achieve superior detection performance but resist straightforward explanation. Regulators must therefore balance predictive power against evidentiary requirements, potentially favoring inherently interpretable architectures such as neural additive models or explanation-augmented gradient boosting over black-box deep learning, even at the cost of some detection performance. A fraud alert that cannot be explained to an enforcement attorney or articulated in an administrative proceeding is of limited operational value.

Second, cross-jurisdictional data sharing is a prerequisite for effective detection, and federated learning offers a promising technical path forward. A fund domiciled in the Cayman Islands, managed from New York, with European investors and Asian prime brokerage relationships generates regulatory filings across multiple jurisdictions, none of which possess a complete picture. The fragmentation documented in Section 6.1.2 is not merely an implementation obstacle but a structural feature of the hedge fund industry that sophisticated fraudsters deliberately exploit through regulatory arbitrage. Federated learning frameworks that enable regulatory agencies to train shared detection models without exchanging raw data can address this fragmentation while preserving the confidentiality constraints that currently prevent cross-border data sharing. The research community has developed the technical foundations for federated fraud detection; what remains is the institutional coordination to deploy these methods operationally.

Key Takeaways for Researchers

Academic researchers in machine learning, finance, and financial regulation should recognize three high-impact opportunities.

First, the creation of benchmark datasets is the most critical enabler for progress. The absence of a public benchmark dataset for hedge fund fraud detection has fragmented the literature, prevented reproducible comparisons, and impeded methodological advancement. Unlike credit card fraud, which benefits from widely used benchmarks, or financial statement fraud, which can draw on publicly available accounting data linked to enforcement actions, hedge fund fraud detection lacks any comparable resource. The two-track approach outlined in Section 6.1.1—combining synthetic data generation with differentially private release of regulatory enforcement data—offers a feasible path forward. Researchers with access to commercial hedge fund databases or regulatory agencies with enforcement-labeled datasets are positioned to make a foundational contribution by constructing and releasing such a benchmark.

Second, multi-modal fusion that integrates returns, filings, networks, and alternative data is underexplored and high-potential. Most existing detection methods operate on a

single data modality: return-based statistical tests analyze time series in isolation, NLP methods process filings without reference to quantitative performance, and network approaches examine relational structures without incorporating fund-level characteristics. Yet the most sophisticated fraud schemes exhibit anomalies across multiple modalities simultaneously—fabricated returns that are statistically implausible, textual descriptions in filings that contradict quantitative factor exposures, and network structures that reveal undisclosed conflicts of interest. Attention-based multi-modal transformers, cross-modal contrastive learning, and hierarchical fusion architectures (Section 6.2.4) represent promising technical directions that have been applied successfully in other domains but remain largely unexplored for hedge fund fraud detection.

Third, adversarial robustness with domain-specific threat models deserves far more attention. The adversarial machine learning literature has focused primarily on image classification, where perturbations are constrained by perceptual similarity, and credit card fraud, where adversaries are unsophisticated. Hedge fund fraud presents a qualitatively different adversarial problem: the adversary is highly quantitative, has access to the same modeling tools as the detector, and operates under economic constraints—such as the requirement to produce returns that satisfy investor expectations and pass administrator review—that differ fundamentally from the ℓ_p-norm perturbation budgets standard in the adversarial ML literature. Certified robustness bounds, game-theoretic adversarial modeling, and red-teaming exercises with domain experts (Section 6.3.1) can provide more realistic assessments of detection robustness than generic adversarial perturbations. This research direction not only advances the technical frontier but also addresses the practical reality that deployment of detection systems creates an arms race in which fraudsters adapt their behavior in response to known detection methods.

The Co-Evolution of Fraud and Detection

The history of financial fraud detection is a history of co-evolution. Statistical anomaly detection led to the engineering of statistically plausible fabricated returns. Benford's law analysis prompted fraudsters to engineer digit distributions that pass Benford tests.

Return smoothing detection motivated the development of more sophisticated NAV manipulation techniques that exhibit plausible serial correlation. This adversarial dynamic will continue as AI methods improve: the deployment of machine learning models for fraud surveillance will, inevitably, be met with adaptive strategies designed to evade those models. The research community must therefore focus not only on detection methods that perform well on historical data but on methods that remain effective under strategic manipulation by sophisticated adversaries. This requires a shift from the standard supervised learning paradigm—in which the data distribution is assumed to be stationary—to an adversarial learning paradigm in which the defender and the fraudster are engaged in a repeated game.

The ten open problems articulated in Section 6 represent not merely technical challenges but institutional ones. Progress demands collaboration among regulatory agencies who possess enforcement-labeled data, academic researchers who develop robust detection methods, and industry practitioners who understand the operational realities of hedge fund management. The creation of benchmark datasets, the deployment of federated learning across jurisdictions, and the evaluation of human-AI collaboration in investigation workflows each require partnerships that transcend traditional academic boundaries. The field is at an inflection point: the technical foundations for AI-based fraud detection have been established, the regulatory imperative is clear, and the economic stakes are enormous. What remains is the coordinated effort to translate promising methods into operationally deployed, robustly tested, and legally compliant systems that can protect investors and preserve market integrity at the scale that the modern hedge fund industry demands.

Reproducibility Statement

This paper is a qualitative survey of AI-based methods for hedge fund fraud detection. No original computational experiments were conducted, and therefore there are no experimental results requiring reproduction. However, to facilitate validation of our literature

synthesis and to support future research building on this work, we provide the following transparency statement regarding our methodology and data sources.

Literature search protocol. The survey followed a systematic search protocol informed by the SALSA methodology (**?**) across five primary scholarly databases: Scopus, Web of Science, IEEE Xplore, the Social Science Research Network (SSRN), and Google Scholar. We employed a structured Boolean search query combining domain terms ("hedge fund," "alternative investment," "investment fund") with fraud-related terms ("fraud," "manipulation," "anomaly") and method terms ("machine learning," "artificial intelligence," "deep learning," "neural network") across titles, abstracts, and keywords. The search period covered publications from 2000 to 2025. The initial database searches yielded approximately 500 potentially relevant publications. Title and abstract screening reduced this set to 120 papers warranting full-text review. After detailed examination against inclusion criteria—peer-reviewed publications or widely cited preprints addressing fraud or anomaly detection in investment funds using AI/ML methods—80 papers met all criteria and form the core of this survey. An additional 25 papers on broader topics (financial regulation, general fraud detection, foundational ML methods) were included for contextual background, bringing the systematically identified corpus to 105 papers. Reference lists of highly cited papers were manually reviewed using snowballing to identify additional relevant studies. The full bibliography includes further foundational and methodological references cited for technical context. We emphasize that our qualitative synthesis focuses on methodological insights and contextual applicability rather than exhaustive enumeration.

Data sources. All datasets referenced in this survey are publicly documented in their originating publications. The primary commercial hedge fund databases cited throughout the paper—Lipper TASS (now Refinitiv), Hedge Fund Research (HFR), BarclayHedge, and Morningstar—require institutional subscriptions and impose licensing restrictions that prohibit redistribution. These databases are, however, widely accessible to academic researchers through university library subscriptions and to practitioners through commercial licenses. Regulatory filing data from the U.S. Securities and Exchange Commission,

including Form ADV, Form D, and Form 13F, are publicly accessible without restriction through the SEC's Electronic Data Gathering, Analysis, and Retrieval (EDGAR) system at `https://www.sec.gov/edgar`. European regulatory data under the Alternative Investment Fund Managers Directive (AIFMD) are available through national competent authorities, though access procedures vary by jurisdiction.

Limitations and constraints. Several cited studies employ proprietary enforcement datasets maintained by regulatory agencies and cannot be independently verified due to confidentiality restrictions. In such cases, we rely on the methodological descriptions provided in the published papers and note where data access limitations prevent full reproducibility. Additionally, some commercial alternative data sources referenced in Section 2.2.3—including satellite imagery, geolocation analytics, and sentiment data—are available only through vendor-specific licenses and cannot be redistributed. We have documented these constraints where they arise and have prioritized discussion of methods that can be evaluated using publicly accessible data wherever possible.

Code and implementation details. As a survey paper, we do not provide original code. However, many of the detection methods reviewed in Section 4 have publicly available implementations. Canonical machine learning methods—random forests, gradient boosting, support vector machines—are implemented in widely used libraries including scikit-learn (Python), caret (R), and XGBoost. Deep learning architectures are available through TensorFlow, PyTorch, and Keras. Graph neural network implementations are provided in PyTorch Geometric and the Deep Graph Library. Where specific papers have released code, we have cited the associated repositories.

Future researchers seeking to build on this work are encouraged to consult the research agenda in Section 6, which identifies the construction of public benchmark datasets (OP1) as the most critical enabler for reproducible progress in this field. Until such benchmarks exist, full end-to-end reproducibility of hedge fund fraud detection research will remain constrained by data access limitations that are endemic to the domain.

References

Aisha Abdallah, Mohd Aizaini Maarof, and Anazida Zainal. Fraud detection system: A survey. *Journal of Network and Computer Applications*, 68:90–113, 2016. doi: 10.1016/j.jnca.2016.04.007.

Vikas Agarwal, Naveen D. Daniel, and Narayan Y. Naik. Do hedge funds manage their reported returns? *Review of Financial Studies*, 24(10):3281–3320, 2011. doi: 10.1093/rfs/hhr058.

Syed Ejaz Ahmed, Risul Islam Rasel, and Carson K. Leung. A survey on AI-based financial fraud detection. *Journal of Big Data*, 11:57, 2024. doi: 10.1186/s40537-024-00899-x.

Adam L. Aiken, Christopher P. Clifford, and Jesse A. Ellis. Out of the dark: Hedge fund reporting biases and commercial databases. *Review of Financial Studies*, 26(1):208–243, 2013. doi: 10.1093/rfs/hhs100.

Dogu Araci. FinBERT: Financial sentiment analysis with pre-trained language models. *arXiv preprint arXiv:1908.10063*, 2019.

George O. Aragon. Share restrictions and asset pricing: Evidence from the hedge fund industry. *Journal of Financial Economics*, 83(1):33–58, 2007. doi: 10.1016/j.jfineco.2005.11.001.

Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador García, Sergio Gil-López, Daniel Molina, Richard Benjamins, Raja Chatila, and Francisco Herrera. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, 58:82–115, 2020. doi: 10.1016/j.inffus.2019.12.012.

Yang Bao, Bin Ke, Bin Li, Y. Julia Yu, and Jie Zhang. Detecting accounting fraud in publicly traded U.S. firms using a machine learning approach. *Journal of Accounting Research*, 58(1):199–235, 2020. doi: 10.1111/1475-679X.12292.

Frank Benford. The law of anomalous numbers. *Proceedings of the American Philosophical Society*, 78(4):551–572, 1938.

Battista Biggio and Fabio Roli. Wild patterns: Ten years after the rise of adversarial machine learning. *Pattern Recognition*, 84:317–331, 2018. doi: 10.1016/j.patcog.2018.07.023.

Nicolas P. B. Bollen and Veronika K. Pool. Suspicious patterns in hedge fund returns and the risk of fraud. *Review of Financial Studies*, 25(9):2673–2702, 2012. doi: 10.1093/rfs/hhs085.

Richard J. Bolton and David J. Hand. Statistical fraud detection: A review. *Statistical Science*, 17(3):235–255, 2002. doi: 10.1214/ss/1042727940.

Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001. doi: 10.1023/A:1010933404324.

Markus M. Breunig, Hans-Peter Kriegel, Raymond T. Ng, and Jörg Sander. LOF: Identifying density-based local outliers. In *Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data*, pages 93–104. ACM, 2000. doi: 10.1145/342009.335388.

Stephen J. Brown, William N. Goetzmann, Bing Liang, and Christopher Schwarz. Mandatory disclosure and operational risk: Evidence from hedge fund registration. *Journal of Finance*, 63(6):2785–2815, 2008. doi: 10.1111/j.1540-6261.2008.01413.x.

Francesco Cartella, Orlando Anunciacao, Yamato Funabiki, Daisuke Yamaguchi, Toru Akishita, and Olivier Elshocht. Adversarial attacks for tabular data: Application to fraud detection and imbalanced data. In *Proceedings of the AAAI Workshop on Adversarial Machine Learning in Real-World Computer Vision Systems*, 2021.

Raghavendra Chalapathy and Sanjay Chawla. Deep learning for anomaly detection: A survey. *arXiv preprint arXiv:1901.03407*, 2019.

Varun Chandola, Arindam Banerjee, and Vipin Kumar. Anomaly detection: A survey. *ACM Computing Surveys*, 41(3):1–58, 2009. doi: 10.1145/1541880.1541882.

Nitesh V. Chawla, Kevin W. Bowyer, Lawrence O. Hall, and W. Philip Kegelmeyer. SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16:321–357, 2002. doi: 10.1613/jair.953.

Tianqi Chen and Carlos Guestrin. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 785–794. ACM, 2016. doi: 10.1145/2939672.2939785.

Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine Learning*, 20 (3):273–297, 1995. doi: 10.1007/BF00994018.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4171–4186. Association for Computational Linguistics, 2019. doi: 10.18653/v1/N19-1423.

Stephen G. Dimmock and William C. Gerken. Predicting fraud by investment managers. *Journal of Financial Economics*, 105(1):153–173, 2012. doi: 10.1016/j.jfineco.2012.01.002.

Georgios Douzas, Fernando Bacao, and Felix Last. Improving imbalanced learning through a heuristic oversampling method based on k-means and SMOTE. *Information Sciences*, 465:1–20, 2018. doi: 10.1016/j.ins.2018.06.056.

Martin Ester, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining*, pages 226–231. AAAI Press, 1996.

European Parliament and Council of the European Union. Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). Regulation, Official Journal of the European Union, 2024.

Financial Stability Board. Artificial intelligence and machine learning in financial services: Market developments and financial stability implications. Report, Financial Stability Board, 2017.

Ugo Fiore, Alfredo De Santis, Francesca Perla, Paolo Zanetti, and Francesco Palmieri. Using generative adversarial networks for improving classification effectiveness in credit card fraud detection. *Information Sciences*, 479:448–455, 2019. doi: 10.1016/j.ins.2017.12.030.

William Fung and David A. Hsieh. The risk in hedge fund strategies: Theory and evidence from trend followers. *Review of Financial Studies*, 14(2):313–341, 2001. doi: 10.1093/rfs/14.2.313.

William Fung and David A. Hsieh. Measurement biases in hedge fund performance data: An update. *Financial Analysts Journal*, 65(3):36–38, 2009. doi: 10.2469/faj.v65.n3.6.

Mila Getmansky, Andrew W. Lo, and Igor Makarov. An econometric model of serial correlation and illiquidity in hedge fund returns. *Journal of Financial Economics*, 74 (3):529–609, 2004. doi: 10.1016/j.jfineco.2004.04.001.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, volume 27, pages 2672–2680. Curran Associates, 2014.

Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *Proceedings of the 3rd International Conference on Learning Representations*, 2015.

Greg N. Gregoriou and François-Serge Lhabitant. Madoff: A riot of red flags. In *EDHEC Risk Institute Working Papers*. 2009.

Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. A survey of methods for explaining black box models. *ACM Computing Surveys*, 51(5):1–42, 2019. doi: 10.1145/3236009.

William L. Hamilton, Rex Ying, and Jure Leskovec. Inductive representation learning on large graphs. In *Advances in Neural Information Processing Systems*, volume 30, pages 1024–1034. Curran Associates, 2017.

Waleed Hilal, S. Andrew Gadsden, and John Yawney. Financial fraud: A review of anomaly detection techniques and recent advances. *Expert Systems with Applications*, 193:116429, 2022. doi: 10.1016/j.eswa.2021.116429.

Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, 9(8):1735–1780, 1997. doi: 10.1162/neco.1997.9.8.1735.

Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. LightGBM: A highly efficient gradient boosting decision tree. In *Advances in Neural Information Processing Systems*, volume 30, pages 3146–3154. Curran Associates, 2017.

Diederik P. Kingma and Max Welling. Auto-encoding variational Bayes. In *Proceedings of the 2nd International Conference on Learning Representations*, 2014.

Thomas N. Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. In *Proceedings of the 5th International Conference on Learning Representations*, 2017.

Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553): 436–444, 2015. doi: 10.1038/nature14539.

Mervyn K. Lewis. New dogs, old tricks. Why do Ponzi schemes succeed? *Accounting Forum*, 36(4):294–309, 2012. doi: 10.1016/j.accfor.2012.02.002.

Fei Tony Liu, Kai Ming Ting, and Zhi-Hua Zhou. Isolation forest. In *Proceedings of the*

8th IEEE International Conference on Data Mining, pages 413–422. IEEE, 2008. doi: 10.1109/ICDM.2008.17.

Yang Liu, Xiang Ao, Zidi Qin, Jianfeng Chi, Jinghua Feng, Hao Yang, and Qing He. Pick and choose: A GNN-based imbalanced learning approach for fraud detection. In *Proceedings of the Web Conference 2021*, pages 3168–3177. ACM, 2021. doi: 10.1145/3442381.3449989.

Andrew W. Lo. Risk management for hedge funds: Introduction and overview. *Financial Analysts Journal*, 57(6):16–33, 2001. doi: 10.2469/faj.v57.n6.2490.

Tim Loughran and Bill McDonald. When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks. *Journal of Finance*, 66(1):35–65, 2011. doi: 10.1111/j.1540-6261.2010.01625.x.

Tim Loughran and Bill McDonald. Textual analysis in accounting and finance: A survey. *Journal of Accounting Research*, 54(4):1187–1230, 2016. doi: 10.1111/1475-679X.12123.

Lefteris Loukas, Manos Fergadiotis, Ilias Chalkidis, Eirini Spyropoulou, Prodromos Malakasiotis, Ion Androutsopoulos, and Georgios Paliouras. SEC-BERT: A domain-specific language model for SEC filings. *arXiv preprint arXiv:2203.06952*, 2022.

Scott M. Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*, volume 30, pages 4765–4774. Curran Associates, 2017.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *Proceedings of the 6th International Conference on Learning Representations*, 2018.

Christoph Molnar. *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable*. Leanpub, 2nd edition, 2020.

Eric W. T. Ngai, Yong Hu, Y. H. Wong, Yijun Chen, and Xin Sun. The application of data mining techniques in financial fraud detection: A classification framework and an academic review of literature. *Decision Support Systems*, 50(3):559–569, 2011. doi: 10.1016/j.dss.2010.08.006.

Mark J. Nigrini. *Benford's Law: Applications for Forensic Accounting, Auditing, and Fraud Detection*. John Wiley & Sons, Hoboken, NJ, 2012. doi: 10.1002/9781119203094.

Guansong Pang, Chunhua Shen, Longbing Cao, and Anton Van Den Hengel. Deep learning for anomaly detection: A review. *ACM Computing Surveys*, 54(2):1–38, 2021. doi: 10.1145/3439950.

Andrew J. Patton, Tarun Ramadorai, and Michael Streatfield. Change you can believe in? Hedge fund data revisions. *Journal of Finance*, 70(3):963–999, 2015. doi: 10.1111/jofi.12240.

Clifton Phua, Vincent Lee, Kate Smith, and Ross Gayler. A comprehensive survey of data mining-based fraud detection research. *arXiv preprint arXiv:1009.6119*, 2010.

Tahereh Pourhabibi, Kok-Leong Ong, Booi H. Kam, and Yee Ling Boo. Fraud detection: A systematic literature review of graph-based anomaly detection approaches. *Decision Support Systems*, 133:113303, 2020. doi: 10.1016/j.dss.2020.113303.

Liudmila Prokhorenkova, Gleb Gusev, Aleksandr Vorobev, Anna Veronika Dorogush, and Andrey Gulin. CatBoost: Unbiased boosting with categorical features. In *Advances in Neural Information Processing Systems*, volume 31, pages 6638–6648. Curran Associates, 2018.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "Why Should I Trust You?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1135–1144. ACM, 2016. doi: 10.1145/2939672.2939778.

Cynthia Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5): 206–215, 2019. doi: 10.1038/s42256-019-0048-x.

René M. Stulz. Hedge funds: Past, present, and future. *Journal of Economic Perspectives*, 21(2):175–194, 2007. doi: 10.1257/jep.21.2.175.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, pages 5998–6008. Curran Associates, 2017.

Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. Graph attention networks. In *Proceedings of the 6th International Conference on Learning Representations*, 2018.

Daixin Wang, Jianbin Lin, Peng Cui, Quanhui Jia, Zhen Wang, Yanming Fang, Quan Yu, Jun Zhou, Shuang Yang, and Yuan Qi. A semi-supervised graph attentive network for financial fraud detection. In *Proceedings of the IEEE International Conference on Data Mining*, pages 598–607. IEEE, 2019. doi: 10.1109/ICDM.2019.00070.

Jarrod West and Maumita Bhattacharya. Intelligent financial fraud detection: A comprehensive review. *Computers & Security*, 57:47–66, 2016. doi: 10.1016/j.cose.2015.09.005.

A Systematic Search Protocol

This appendix documents the systematic search methodology employed to identify relevant literature on AI-based detection of hedge fund fraud. The search framework was informed by the SALSA methodology (**?**)—Search, AppraisaL, Synthesis, Analysis—adapted from systematic reviews in clinical research to accommodate the interdisciplinary nature of AI/ML applications in financial fraud detection.

Search Framework and Databases

The systematic search was conducted across five major academic databases to ensure comprehensive coverage of both computer science and finance literature: Scopus, Web of Science, IEEE Xplore (for AI/ML methodology), SSRN (for finance working papers), and Google Scholar (for supplementary coverage and citation tracking). The search period covered publications from 2000 to 2025, with particular emphasis on the period 2015–2025 when deep learning and modern AI techniques became prominent in financial applications.

Query Strings and Search Strategy

The primary search query combined three conceptual blocks using Boolean operators:

- **Financial domain:** ("hedge fund" OR "investment fund" OR "alternative investment")
- **Fraud-related terms:** ("fraud" OR "manipulation" OR "anomaly")
- **AI/ML methods:** ("machine learning" OR "artificial intelligence" OR "deep learning" OR "neural network")

A secondary, broader query was employed to capture related research that might use different terminology: ("fund" OR "portfolio") AND ("fraud detection" OR "anomaly detection") AND ("classification" OR "prediction"). This secondary search helped identify relevant studies that focused on mutual funds or general portfolio fraud rather than exclusively hedge funds.

Inclusion and Exclusion Criteria

Studies were included if they met the following criteria: (1) directly addressed fraud, manipulation, or anomaly detection in investment funds or alternative investments; (2) employed or substantively discussed AI/ML methods for detection, prediction, or classification; (3) were peer-reviewed publications or widely cited preprints with at least 20 citations; and (4) were published in English.

Exclusion criteria eliminated: (1) pure methodology papers without financial application or validation; (2) studies focused exclusively on credit card, payment, or transaction-level fraud without fund-level components; and (3) duplicate publications reporting the same study across multiple venues.

Screening Process and Results

The initial database searches yielded approximately 500 potentially relevant publications. Title and abstract screening reduced this set to 120 papers warranting full-text review. After detailed examination, 80 papers met all inclusion criteria and form the core of this survey. An additional 25 papers on broader topics (financial regulation, general fraud detection, or foundational ML methods) were included for contextual background, bringing the total reference count to 105.

Two independent reviewers conducted the screening process, with disagreements resolved through discussion and consultation with a third reviewer. The reference lists of highly cited papers were manually reviewed using a snowballing technique to identify additional relevant studies not captured by the database searches. This systematic approach ensures that the survey provides comprehensive coverage of the AI-based hedge fund fraud detection literature while maintaining rigorous inclusion standards.

B Feature Engineering Details

This appendix provides detailed specifications for the key features employed in hedge fund fraud detection models. Feature engineering represents a critical component of detection

systems, as the quality and informativeness of input features directly determine model performance. The features are organized by category, with mathematical definitions, data sources, and representative citations.

Overview

Effective fraud detection requires features that capture different aspects of fund behavior: statistical properties of returns, adherence to mathematical laws, textual characteristics of disclosures, network relationships, and temporal patterns. Table 4 summarizes the most commonly used features across the literature, organized by category.

Data Sources and Construction

Statistical features are constructed from monthly return time series, typically requiring at least 24–36 months of data for reliable estimation. Benford's law features can be computed from returns, NAV values, or fee disclosures. Textual features are extracted from regulatory filings (Form ADV in the U.S., AIFMD disclosures in the EU), annual letters, and offering documents. Network features require relationship data from regulatory databases, commercial data vendors, or manual extraction from disclosures.

The construction of robust features must address several challenges. Return-based features are susceptible to backfill bias and survivorship bias in commercial databases. Textual features require careful preprocessing (tokenization, stopword removal, domain-specific term handling). Network features are often incomplete due to selective disclosure and evolving relationship structures. Temporal features must be computed on rolling windows to enable real-time detection while maintaining sufficient statistical power.

Feature Selection and Dimensionality Reduction

Given the high dimensionality of feature spaces (often 50–200 features), most studies employ feature selection techniques. Common approaches include LASSO regularization for linear models, tree-based feature importance for ensemble methods, and mutual information criteria for neural networks. Dimensionality reduction via PCA or autoencoders is

less common due to the loss of interpretability, which is critical for regulatory applications and explainability requirements.

C Glossary of Terms

[style=nextline]

Adversarial Training A machine learning technique where models are trained on both genuine and intentionally perturbed examples to improve robustness against adversarial attacks.

AIFMD (Alternative Investment Fund Managers Directive) European Union regulation requiring registration, disclosure, and oversight of alternative investment fund managers including hedge funds.

AUC (Area Under the Curve) A performance metric for classification models measuring the area under the ROC curve; ranges from 0 to 1, with 0.5 indicating random performance and 1.0 perfect classification.

Autoencoder A neural network architecture that learns compressed representations by training to reconstruct its input, commonly used for anomaly detection by identifying samples with high reconstruction error.

Backfill Bias The artificial inflation of historical performance when funds selectively report past returns to databases only after establishing a track record.

Benford's Law A mathematical principle stating that in many naturally occurring datasets, leading digits follow a logarithmic distribution with "1" appearing as the first digit about 30% of the time; deviations suggest manipulation.

BERT (Bidirectional Encoder Representations from Transformers) A transformer-based language model that generates contextualized word embeddings; FinBERT is a variant fine-tuned on financial text.

Class Imbalance A dataset characteristic where one class (e.g., fraud cases) is substantially less frequent than others, requiring specialized sampling or algorithmic techniques for effective learning.

Concept Drift The phenomenon where the statistical properties of the target variable change over time, requiring model retraining or adaptive learning strategies.

DBSCAN (Density-Based Spatial Clustering of Applications with Noise) An unsupervised clustering algorithm that groups densely packed points and identifies outliers as potential anomalies.

Dodd-Frank Act U.S. financial reform legislation enacted in 2010 requiring hedge fund registration with the SEC and periodic disclosure of positions and risk metrics.

Ensemble Methods Machine learning techniques that combine multiple base models (e.g., decision trees) to improve predictive performance and robustness; examples include Random Forest and XGBoost.

Evasion Attack An adversarial attack where fraudsters intentionally modify their behavior to avoid detection by a known fraud detection model.

F1 Score The harmonic mean of precision and recall: $F_1 = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$; particularly useful for imbalanced datasets.

Form ADV The SEC registration form for investment advisers in the United States, containing information about business practices, fee structures, conflicts of interest, and disciplinary history.

GAT (Graph Attention Network) A graph neural network architecture that learns importance weights for neighboring nodes, enabling selective aggregation of information in network-structured data.

GCN (Graph Convolutional Network) A neural network architecture that operates on graph-structured data by aggregating features from neighboring nodes through convolutional operations.

GNN (Graph Neural Network) A class of deep learning models designed to process graph-structured data by learning representations that capture both node features and topological structure.

Gradient Boosting An ensemble learning method that builds models sequentially, with each new model correcting errors of the previous ones; XGBoost and LightGBM are popular implementations.

Hedge Fund A pooled investment vehicle that employs diverse strategies (long/short equity, global macro, event-driven) and typically has limited regulatory oversight and investor restrictions.

Isolation Forest An anomaly detection algorithm that identifies outliers by measuring how easily samples can be isolated in randomly constructed decision trees.

LIME (Local Interpretable Model-Agnostic Explanations) An explainability technique that approximates complex model predictions locally using interpretable linear models.

LSTM (Long Short-Term Memory) A recurrent neural network architecture with gating mechanisms that can learn long-range temporal dependencies, often used for time series analysis.

NAV (Net Asset Value) The total value of a fund's assets minus liabilities, typically calculated per share; the primary metric reported to investors.

Ponzi Scheme A fraudulent investment operation that pays returns to earlier investors using capital from new investors rather than from legitimate profits.

Random Forest An ensemble learning method that constructs multiple decision trees during training and outputs the mode (classification) or mean (regression) of individual tree predictions.

Serial Correlation The correlation between time-series observations at different time lags; abnormally high serial correlation in hedge fund returns may indicate return

smoothing.

SHAP (SHapley Additive exPlanations) An explainability framework based on cooperative game theory that assigns each feature an importance value for a particular prediction.

SupTech (Supervisory Technology) Technology-based solutions employed by financial regulators for data collection, risk assessment, and market surveillance.

Survivorship Bias The distortion of performance statistics when failed or closed funds are excluded from databases, leading to overestimation of average returns.

XGBoost (eXtreme Gradient Boosting) An optimized implementation of gradient boosting that uses regularization, parallel processing, and efficient tree construction; widely used in fraud detection competitions.

Table 4: Key Features in Hedge Fund Fraud Detection

Feature	**Formula/Description**	**Category**	**Reference**
First-order autocorrelation	$\rho_1 = \text{Corr}(r_t, r_{t-1})$	Statistical	Getmansky et al. (2004)
Sharpe ratio	$SR = \bar{r}/\sigma_r$	Statistical	?
Maximum drawdown	$MDD = \max_t \left(\frac{\max_{s \le t} P_s - P_t}{\max_{s \le t} P_s} \right)$	Statistical	?
Kurtosis	Excess kurtosis: $\text{Kurt}(r) - 3$	Statistical	?
Discontinuity at zero	Kink in return distribution at zero (Bollen-Pool measure)	Statistical	?
Hurst exponent	H estimated via rescaled range (R/S) analysis	Statistical	?
First-digit test	$\chi^2 = \sum_{d=1}^{9} \frac{(O_d - E_d)^2}{E_d}$ for Benford's law	Benford	?
Second-digit test	Benford's law applied to second significant digit	Benford	?
Summation test	Cumulative conformity to Benford across digit positions	Benford	?
Fog index	Readability: $0.4[(w/s) + 100(c/w)]$	Textual	?
FinBERT sentiment	BERT-based financial sentiment score $\in [-1, 1]$	Textual	?
Boilerplate deviation	$1 - \text{CosSim}(\text{doc}, \text{template})$	Textual	?
Degree centrality	Number of connections in fund-service-provider network	Network	?
Betweenness centrality	$\sum_{s \neq v \neq t} \frac{\sigma_{st}(v)}{\sigma_{st}}$	Network	?
Related-party count	Number of related-party relationships disclosed	Network	?
Regime indicator	Binary indicator from Hidden Markov Model state	Temporal	?
Change-point score	Bayesian change-point detection probability	Temporal	?