A Unified Detection Pipeline Framework (C1)

Section 3 – AI-Based Detection of Hedge Fund Fraud

Joerg Osterrieder

Zurich University of Applied Sciences (ZHAW)

2025

Outline

1. Pipeline Overview: Five Stages
2. Stage 1: Data Ingestion and Integration
   - Temporal alignment, data quality, entity resolution, fusion
3. Stage 2: Feature Engineering (5 families)
   - Statistical, Benford's law, textual, network, temporal
4. Stage 3: Model Selection and Training (6 families)
   - Classical ML, deep learning, anomaly detection, NLP, GNN, generative
5. Stage 4: Explainability and Interpretation
6. Stage 5: Deployment and Monitoring
7. Summary

Pipeline Overview: Five Stages

1. **Data Ingestion & Integration**
   Multi-source collection, temporal alignment, entity resolution
2. **Feature Engineering**
   5 feature families: statistical, Benford, textual, network, temporal
3. **Model Selection & Training**
   6 method families matched to fraud types
4. **Explainability & Interpretation**
   Regulatory transparency (EU AI Act Art. 13)
5. **Deployment & Monitoring**
   Batch/real-time, drift detection, human-in-the-loop

Key Design Principles

- Pipeline is **not strictly unidirectional**
- Feedback loop: deployment $\rightarrow$ feature engineering, model training
- Incorporates investigator judgments, adapts to evolving fraud patterns
- No prior work assembles these components into a coherent framework *tailored to hedge funds*

Three Purposes

- Researchers: structured lens for positioning contributions
- Practitioners: engineering blueprint for operational systems
- Gap analysis: reveals methodological gaps $\rightarrow$ research agenda

Source: Paper Section 3.1 – Contribution C1

Stage 1: Data Ingestion and Integration

- Assemble a coherent analytical dataset from sources differing in **structure, frequency, reliability, provenance**
- Four layers: returns (monthly, numerical), filings (quarterly/annual, text), alternative (continuous, heterogeneous), synthetic (on-demand)
- Four sub-problems must be solved:

Temporal Alignment

Different cadences: returns (monthly), filings (quarterly), alt data (continuous)

Data Quality

Survivorship (+242 bp), backfill (+442 bp), selection biases

Entity Resolution

Same fund → different IDs across databases (TASS, HFR, CRD, LEI)

Multi-Source Fusion

Early vs. late fusion architectures

Source: Paper Section 3.2

Stage 1: Temporal Alignment Challenge

Reporting Cadences

- Return data: monthly, 30–60 day lag
- Form ADV: annual + ad hoc amendments
- Form 13F: quarterly, 45-day delay
- News / social media: continuous, irregular

Alignment Strategies

- Aggregate to quarter-end $\Rightarrow$ discards intra-quarter dynamics
- Multi-resolution architectures: process each stream at native frequency before fusion

Why It Matters

- End-of-quarter return spikes (Bollen & Pool, 2012) = potential manipulation signals
- Choice of alignment strategy determines which temporal patterns remain visible
- Misalignment can introduce spurious correlations or mask genuine signals

Practical recommendation: hybrid architecture that preserves native frequency per source, fuses at the model input stage

Source: Bollen & Pool (2012); paper Section 3.2

Stage 1: Data Quality and Bias Correction

- Biases propagate through the **entire pipeline** if uncorrected

Bias Correction Procedures

- **Backfill**: restrict analysis to post-reporting-inception returns only
- **Survivorship**: require databases with graveyard (defunct) fund records
- **Selection**: model the reporting decision process itself as an informative signal

Fraud-Specific Asymmetry

- Detected frauds $\rightarrow$ graveyard section
- Undetected frauds remain in live data
- "Clean" training class is contaminated
- Models trained on biased data underestimate base rate of fraud
- May miss signatures of currently active schemes

Source: Fung & Hsieh (2009); Agarwal et al. (2011); paper Section 3.2

Stage 1: Entity Resolution

The Problem

- Same fund $\Rightarrow$ different identifiers across sources:
    - TASS fund ID $\neq$ HFR ID $\neq$ SEC CRD number $\neq$ LEI
- Managers may **deliberately obscure** connections to prior failed/sanctioned funds
- Inability to link is itself a **fraud-relevant signal**

Methods

- Approximate string matching (fund/manager names)
- Shared-attribute clustering (addresses, auditors, administrators, prime brokers)
- Graph-based entity resolution (propagate identity evidence through shared relationships)

Post-Dodd-Frank Improvement

- Form ADV filing since 2010 provides **stable CRD numbers** for US-registered advisers
- CRD serves as linkage key across databases
- Substantially improved entity resolution for US funds

Remaining Gaps

- Non-US funds lack stable identifiers
- Offshore structures (Cayman Islands, BVI) deliberately fragment entity trails
- Graph-based methods offer robustness but at higher computational cost

---

Source: Brown et al. (2008); paper Section 3.2

Stage 2: Feature Engineering – Five Families

Family	Key Features	Target Fraud Types	Data Source
Statistical	ρ_1, Sharpe, skewness, kurtosis, max drawdown, Hurst H	Performance fabrication, NAV manipulation	Return series
Benford's law	First-digit χ^2, second-digit, summation test, KS statistic	Data fabrication	Returns, NAVs
Textual	Fog index, FinBERT sentiment, boilerplate deviation, topic drift	Strategy misrep., regulatory fraud	Form ADV, letters
Network	Auditor risk, manager history, co-investment centrality, clustering	Allocation fraud, serial offenders	Filings, databases
Temporal	HMM regimes, change-points, calendar effects, momentum/reversal	All types (dynamic dimension)	Return series

Source: Paper Section 3.3

Statistical Features

Serial Correlation

- $\rho_1 = \mathrm{Corr}(r_t, r_{t-1})$: proxy for return smoothing
- Typical $\rho_1 = 0.3$–0.5 for illiquid positions
- Higher-order: ρ_2, ρ_3, Ljung-Box Q-statistic
- Abnormally high for funds claiming liquid assets $\Rightarrow$ NAV manipulation

Distributional Discontinuity

- Bollen-Pool "kink" at zero: excess small positives, deficit of small negatives
- Quantified via kernel density estimation or histogram structural break

Higher Moments & Risk-Adjusted

- Sharpe ratio $S = \bar{r}/\sigma_r$: implausibly high $\Rightarrow$ fabrication
- Skewness, excess kurtosis: fabricated returns $\approx$ zero skew, low kurtosis
- Maximum drawdown: difficult to fake over long horizons

Long-Memory Detection

- Hurst exponent H (R/S analysis, DFA)
- $H \gg 0.5$: persistent serial dependence $\Rightarrow$ smoothing
- Combined with GARCH residual analysis for volatility clustering

Source: Getmansky et al. (2004); Lo (2001); paper Section 3.3.1

Benford's Law Features

Foundation

- Leading digit d: $P(d) = \log_{10}(1 + 1/d)$
- Naturally occurring data follow this logarithmic distribution
- Human-generated / engineered numbers often fail to reproduce it

Three Test Types

- **First-digit test**: $\chi^2 = \sum_{d=1}^{9} \frac{(O_d - E_d)^2}{E_d}$
- **Second-digit test**: more sensitive to subtle manipulation (fraudsters engineer first digits but neglect second)
- **Summation test**: detects round-number manipulation in reported amounts

ML Feature Space

- 9 first-digit frequencies + 10 second-digit frequencies + 2 test statistics (χ^2, KS)
- **21-dimensional feature vector** per fund
- Enables detection of complex digit manipulation patterns:
  - E.g., first digits conform but second digits anomalously uniform

Limitations

- Low statistical power for short return histories (< 60 months)
- Knowledgeable fraudster can engineer conformity
- Most valuable as one component in multi-feature pipeline

Source: Benford (1938); Nigrini (2012); paper Section 3.3.2

Textual Features from Regulatory Filings

Filing Complexity / Readability

- Gunning Fog: $0.4 \times (\text{ASL} + \text{PHW})$
- Firms engaged in misconduct tend to produce more complex filings
- Word count, sentence count, type-token ratio

Sentiment Analysis

- FinBERT: BERT fine-tuned on financial text
- SEC-BERT: pre-trained on EDGAR filings
- Aggregate sentiment, sentiment volatility, proportion of hedging language

Boilerplate Deviation

- Cosine similarity vs. peer-template mean TF representation
- Unusual deviation (vague *or* suspiciously precise) $\Rightarrow$ scrutiny
- **Temporal trajectory** more informative than single snapshot:
  - Deteriorating readability
  - Increasing hedging language
  - Growing divergence from prior filings

Topic Modeling

- LDA / neural topic models on strategy descriptions
- Cross-modal consistency: text strategy vs. quantitative factor exposures

Source: Araci (2019); Loukas et al. (2022); paper Section 3.3.3

Network and Relational Features

Fund–Service-Provider Graphs

- Bipartite graph: funds $\leftrightarrow$ auditors, administrators, custodians
- Small, non-Big-Four auditors over-represented among sanctioned funds
- Auditor *change* from reputable to small firm + other risk signals $\Rightarrow$ weakening oversight
- Node features: client count, historical sanctions, change frequency

Manager History Networks

- "Serial offenders": new fund after prior failure/sanction
- Prior sanctions = significant fraud predictor (Dimmock & Gerken, 2012)

Co-Investment / Capital Flow Networks

- Funds sharing common investors or correlated capital flows
- Centrality measures as fund-level features:
  - **Betweenness**: bridges disconnected investor communities (Ponzi signature)
  - **Degree**: number of connections
  - **Eigenvector**: importance of neighbors
  - **Clustering coefficient**: local density

"Guilt by association": individually inconspicuous funds embedded in suspicious relational structures

Source: Brown et al. (2008); Dimmock & Gerken (2012); paper Section 3.3.4

Temporal Features

Regime Detection (HMMs)

- Hidden Markov Models: "normal" vs. "manipulated" regime
- Different mean, variance, autocorrelation per state
- Transition probabilities and regime timing as features
- Distinguishes legitimate adaptation from suspicious behavioral shifts

Change-Point Detection

- Patton & Ramadorai (2015): structural breaks precede fund failure
- Bayesian Online Change Point Detection (BOCPD)
- Features: number of change points, spacing, magnitude
- Frequent, large shifts without identifiable market events $\Rightarrow$ suspicious

Calendar Effects

- Consistently higher December returns
- Abnormally positive last-trading-day-of-quarter returns
- Signals end-of-period NAV manipulation
- Computed as month/day-of-quarter dummy coefficients

Momentum / Reversal Patterns

- Autocorrelation at multiple lags
- Legitimate strategies: characteristic momentum-reversal from investment style
- Fabricated returns: artificially smooth momentum *without* mean-reversion imposed by fundamentals

Source: Patton & Ramadorai (2015); paper Section 3.3.5

Feature Families: Visual Summary

trix or radar showing five feature families (statistical, Benford, textual, network, temporal) mapped to five fraud types with relative imp

Source: Paper Section 3.3

Stage 3: Model Selection – Six Families

Family	Key Methods	Strengths	Best For
Classical ML	Logistic regression, SVM, RF, XGBoost	Interpretable, handles mixed features	Tabular, small n
Deep learning	LSTM, CNN, Transformer	Temporal patterns, nonlinear	Sequential data
Anomaly detection	Isolation Forest, LOF, autoencoder	No labels needed	Label-scarce
NLP / text mining	FinBERT, SEC-BERT, TF-IDF	Text signals, cross-modal	Filings
Graph neural networks	GCN, GAT, GraphSAGE	Relational "guilt by association"	Network data
Generative	GANs, VAEs	Anomaly detection + augmentation	Class imbalance

Context-specific challenges: extreme class imbalance, small sample sizes, heterogeneous fraud types, multi-modal inputs, adversarial dynamics.

Source: Paper Section 3.4

Classical ML: Logistic Regression, SVM, RF, XGBoost

Logistic Regression

- Interpretable baseline: coefficients = log-odds contributions
- Dimmock & Gerken (2012) on Form ADV: $\mathrm{AUC} \approx 0.65$–$0.70$
- Limitation: assumes linear feature-response

SVM

- Maximum-margin hyperplanes
- One-class SVM: semi-supervised, no fraud labels needed
- Good on small datasets; limited scalability and interpretability

Random Forests

- Hundreds of trees, bootstrap + random features
- Handles high-dimensional, mixed-type features naturally
- Robust to outliers, missing values
- Feature importance via permutation / mean decrease in impurity

Gradient Boosting (XGBoost, LightGBM, CatBoost)

- State of the art for tabular classification
- Built-in class imbalance handling
- Stacking ensembles: $F_1 \sim$**0.88** (Hilal et al., 2022)
- CatBoost: native categorical feature support

Source: Breiman (2001); Chen et al. (2016); Hilal et al. (2022); paper Section 3.4.1

Deep Learning: LSTM, CNN, Transformer

LSTM

- Gated memory cells capture long-range sequential dependencies
- Natural fit for monthly return time series
- Learns temporal patterns: gradual smoothing onset, regime transitions

CNN

- Returns $\to$ 2D representation (Gramian Angular Fields, recurrence plots)
- Spatial pattern detection on visual representations
- "Smooth upward ramp" of Ponzi scheme

Transformer

- Self-attention over all sequence positions
- Handles long-range dependencies in sparse monthly series
- Attention weights = built-in explainability

Key challenge: hedge fund return series are extremely short (60–120 months) and < 100 positive labels exist. Regularization, pre-training, and transfer learning are **essential** to prevent overfitting.

Source: Hochreiter & Schmidhuber (1997); Vaswani et al. (2017); paper Section 3.4.2

Anomaly Detection: Isolation Forest, LOF, Autoencoders

Isolation Forest

- Random recursive partitioning; anomalies isolated with fewer splits
- Efficient, high-dimensional, no distributional assumptions
- Detects *global* anomalies (vs. entire population)

Local Outlier Factor (LOF)

- Local density deviation vs. k-nearest neighbors
- Detects *local* anomalies (unusual within strategy peer group)
- Critical: strategy-specific norms differ substantially

Deep Autoencoders

- Learn compressed representation of normal behavior
- High reconstruction error $\Rightarrow$ anomaly
- $\mathrm{AUC} \approx$ **0.79** on hedge fund returns (Chalapathy & Chawla, 2019)
- Competitive with supervised methods *without requiring fraud labels*

DBSCAN

- Density-based clustering; noise points $=$ potential anomalies
- Fund not assignable to any peer group $\Rightarrow$ possible strategy misrepresentation

Unsupervised methods are especially valuable when confirmed fraud labels are scarce and potentially biased toward historically detected types.

Source: Liu et al. (2008); Breunig et al. (2000); Chalapathy & Chawla (2019); paper Section 3.4.3

NLP and Graph Neural Network Methods

NLP / Text Mining

- Evolution: bag-of-words → TF-IDF → word2vec → transformers
- **FinBERT**: financial-domain sentiment
- **SEC-BERT**: EDGAR-specific pre-training
- Applications:
  - Detect vague/evasive strategy descriptions
  - Filing inconsistencies across time
  - Text-quant divergence (stated strategy vs. factor exposures)
- Multi-modal fusion (NLP + returns) more robust: fraudster must maintain both textual *and* statistical consistency

Graph Neural Networks

- **GCN**: spectral graph convolutions, neighborhood aggregation
- **GAT**: attention-weighted neighbor contributions (heterogeneous graphs)
- **GraphSAGE**: inductive learning on unseen nodes (new fund cold-start)
- **Temporal knowledge graphs**: time-stamped edges capture "flight from oversight" patterns

Key Advantage

- "Guilt by association": individually benign funds in suspicious relational structures

Key Limitation

- Graph construction requires entity resolution across databases

Source: Kipf & Welling (2017); Velickovic et al. (2018); paper Sections 3.4.4–3.4.5

Generative and Synthetic Methods

Dual Role

- *Anomaly detection*: learn normal distribution, flag deviations
- *Data augmentation*: generate synthetic fraud examples for class imbalance

GAN-Based Anomaly Detection

- BiGAN, AnoGAN: generator learns normal returns
- High reconstruction error in latent space $\Rightarrow$ anomalous
- Adversarial training captures non-Gaussian features and temporal dependencies

Synthetic Data Generation

- Beyond SMOTE: conditional GANs and VAEs
- Conditional generation essential: different fraud types $\rightarrow$ different signatures
- Wasserstein GAN: earth-mover distance for financial time series

Validation Requirements

- Generated samples must pass domain-specific plausibility checks:
  - Realistic return magnitudes
  - Appropriate serial correlation
  - Sensible Sharpe ratios

Source: Goodfellow et al. (2014); Kingma & Welling (2014); paper Section 3.4.6

Model Families: Visual Summary

ilies – Comparison of six model families showing reported performance ranges (AUC, F1), data requirements, interpretability level, and

Source: Paper Section 3.4

Stage 4: Explainability and Interpretation

SHAP Values

- Decompose each prediction into per-feature contributions (Shapley values)
- Example: "35% from high ρ_1, 25% from non-Big-Four auditor with 2 prior sanctions, 20% from deteriorating readability, 20% from peripheral network position"
- Maps directly to SEC investigative categories

LIME

- Local interpretable model around each prediction
- Model-agnostic: works with deep nets, GNNs
- Tailored case-by-case explanations

Attention Visualization

- Transformer: which historical periods are most relevant
- GAT: which relational connections drove the score
- Intrinsic explainability (no post-hoc layer needed)

Regulatory Requirement

- EU AI Act Art. 13: "sufficiently transparent to enable deployers to interpret"
- Opaque probability scores → insufficient
- Structured explanation → actionable investigation plan
- Three independent lines of inquiry per alert: return analysis, operational due diligence, filing review

Source: Lundberg & Lee (2017); Ribeiro et al. (2016); EU AI Act Art. 13; paper Section 3.5

Stage 5: Deployment and Monitoring

Processing Cadence

- Hedge fund surveillance is **batch-oriented** (monthly/quarterly data)
- Hybrid: batch scoring of full fund universe + event-triggered inter-batch reassessment

Concept Drift Detection

- Models degrade: legitimate strategies evolve; fraudsters adapt
- ADWIN: variable-length window, distribution comparison
- DDM: monitors error rate vs. historical baseline
- Hybrid retraining: scheduled + drift-triggered emergency updates

Human-in-the-Loop

- Designed for **collaboration, not replacement**
- Investigator feedback loop $\rightarrow$ feature engineering and model training
- Active learning: query cases most likely to improve decision boundary
- Reduces alert fatigue from false positives
- EU AI Act Art. 14: human oversight requirement satisfied

Alert Prioritization

- Composite score: fraud probability $\times$ AUM exposure $\times$ novelty $\times$ tractability
- Transforms raw flags into manageable investigation queue

Source: Pang et al. (2021); EU AI Act Art. 14; paper Section 3.6

ure – Full five-stage pipeline diagram with forward data flow, feedback loops (investigator feedback, drift-triggered retraining), and frau

Source: Paper Figure 1; Section 3

Summary: The Detection Pipeline (C1)

1. The **five-stage pipeline** (ingestion $\rightarrow$ features $\rightarrow$ models $\rightarrow$ explainability $\rightarrow$ deployment) provides the first hedge-fund-specific end-to-end framework
2. **Data ingestion** must solve temporal alignment, bias correction, entity resolution, and multi-source fusion
3. **Five feature families** (statistical, Benford, textual, network, temporal) capture complementary fraud signals across different data modalities
4. **Six model families** are matched to specific fraud types and data characteristics; tree-based ensembles currently dominate tabular detection ($F_1 \sim 0.88$)
5. **Explainability** is not optional: EU AI Act mandates transparency for high-risk AI; SHAP/LIME/attention provide structured explanations
6. **Deployment** requires batch+event processing, drift detection, human-in-the-loop feedback, and GRC integration
7. The pipeline is **not unidirectional**: feedback from deployment enables continuous adaptation

Source: Paper Section 3 – Contribution C1