# Review of AI-Based Detection Methods

## Section 4 – AI-Based Detection of Hedge Fund Fraud

Joerg Osterrieder

Zurich University of Applied Sciences (ZHAW)

2025

# Outline

# Classical Statistical and Rule-Based Approaches

**Benford's Law (Nigrini, 2012)**

- Applied retrospectively to Madoff: anomalies in **9/10 tests**
- Fabricators are poor intuitive generators of logarithmic digit distributions
- Requires large samples; limited power for $<$ 120–240 monthly observations
- Aware fraudsters can engineer conformity

**Serial Correlation (Getmansky et al., 2004)**

- MA($k$) model: smoothing component from managed pricing
- 30–40% of TASS funds show significant positive $\rho_1$
- Smoothing parameters now standard ML features

**Bollen & Pool (2012) – Distributional "Kink"**

- Discontinuity at zero: excess small positives, deficit of small negatives
- Correctly identified ∼**50%** of funds subsequently facing SEC enforcement

**Operational Risk / Filing-Based**

- Brown et al. (2008): **omega-score** from Form ADV data – governance and organizational risk
- Dimmock & Gerken (2012): logistic regression on SEC filing data – past violations, ownership, custody $\Rightarrow \mathrm{AUC} \approx$ 0.65–0.70

**Collective limitation:** each method detects one specific signature of one fraud type. High false positive rates when deployed independently.

**Source: Paper Section 4.1**

# Tree-Based and Ensemble Methods

**Random Forests**

- Aggregate hundreds of independently grown trees
- Robust to overfitting, tolerant of missing data
- Handle mixed feature types (numerical + categorical) without standardization

**Gradient Boosting**

- XGBoost, LightGBM, CatBoost
- Sequential tree construction correcting residuals
- Built-in class imbalance handling: weighting, stratified subsampling, cost-sensitive learning

**Key Results**

- **Bao et al. (2020)**: RUSBoost on 28,000+ firm-years linked to AAERs
  - $\mathrm{AUC} = \mathbf{0.725}$, outperforming logistic regression
  - Accounting fraud, but methodology transferable
- **Stacking ensembles** (Hilal et al., 2022):
  - XGBoost + LightGBM + CatBoost via meta-learner
  - $F_1 \approx \mathbf{0.88}$ – highest among individual method families

**Principal Limitation**

- Relies on supervised training: only 50–100 confirmed hedge fund fraud cases
- Positive class is *heterogeneous*: Ponzi $\neq$ NAV manipulation $\neq$ style drift

**Source: Bao et al. (2020); Hilal et al. (2022); paper Section 4.2**

# Why Trees Dominate Hedge Fund Detection

**SHAP Compatibility**

- Exact SHAP computation for tree-based models (Lundberg & Lee, 2017)
- Per-feature attribution: "flagged because $\rho_1$ is $2.3\sigma$ above peer median, auditor has 2 prior sanctions, readability deteriorating"
- Maps directly to SEC investigative categories
- Satisfies EU AI Act transparency requirements

**Class Imbalance Handling**

- Native mechanisms: sample weighting, cost-sensitive learning
- RUSBoost: random undersampling + AdaBoost

**Mixed Feature Types**

- Numerical (return statistics) + categorical (strategy class, auditor ID, jurisdiction)
- No feature standardization required
- CatBoost: native categorical support

**Robustness**

- Tolerant of outliers and missing values
- Stable performance across varying hyperparameters (vs. deep learning sensitivity)
- Can ingest full concatenated 5-family feature vector without dimensionality reduction

Trees provide the best trade-off between **performance, interpretability, and practical robustness** for current hedge fund data conditions.

**Source: Lundberg & Lee (2017); paper Section 4.2**

# Deep Learning Approaches

**LSTM Networks**

- Gated memory cells: long-range sequential dependencies
- Detect gradual shifts in serial correlation, regime-dependent anomalies
- Fraud escalates over months/years: sequential nature matches

**CNN via Gramian Angular Fields**

- Encode returns as 2D images (Gramian, recurrence plots)
- Spatial pattern recognition on visual representations
- Hybrid CNN-LSTM: local features + temporal aggregation

**Transformers**

- Self-attention across arbitrary sequence positions
- Long-range dependencies without vanishing gradients
- Link suspicious patterns across years of sparse monthly data
- Attention weights $\Rightarrow$ intrinsic explainability

**Autoencoders**

- Most directly applicable under label scarcity
- Trained on normal behavior; high reconstruction error = anomaly
- $\mathrm{AUC} \approx$ **0.79** on hedge fund returns (Chalapathy & Chawla, 2019)
- No fraud labels needed; sidesteps data poisoning vulnerability

**Source: Hochreiter & Schmidhuber (1997); Chalapathy & Chawla (2019); paper Section 4.3**

# Deep Learning Limitations in Hedge Fund Context

**Data Scarcity**

- Modern architectures need thousands to millions of examples
- Hedge fund universe: $\sim$10,000–15,000 funds
- At most 120–240 monthly observations per fund
- Fewer than 100 positive (fraud) examples

**Overfitting Risk**

- Models with millions of parameters trained on $< 100$ positives
- May memorize specific fraud fingerprints rather than learn generalizable patterns
- Common practice: single train-test split (inadequate)

**Opacity / Explainability**

- Predictions resist human interpretation
- EU AI Act mandates transparency for high-risk AI
- Post-hoc methods (SHAP, LIME) add complexity but do not fully resolve

**Hyperparameter Sensitivity**

- Small changes in architecture, learning rate, regularization $\Rightarrow$ qualitatively different results
- Undermines reproducibility and reliability of performance claims

**Verdict:** deep learning has theoretical appeal but faces severe practical constraints in the current data regime. Best used in combination with tree-based methods or as anomaly detectors (autoencoders).

**Source: Paper Section 4.3**

# NLP for Financial Filings

### Domain-Specific Lexicons

- Loughran & McDonald (2011): general sentiment lexicons perform poorly on financial text
- "Liability," "tax," "depreciation" = negative in general, neutral in finance
- Financial-domain dictionary: more accurate sentiment classification

### Transformer Models

- **FinBERT**: 87% accuracy on financial sentiment
- **SEC-BERT**: pre-trained on EDGAR filings; improved NER and document classification for regulatory text

### Detection Applications

- Vague/evasive language in strategy descriptions
- Changes in filing complexity over time $\rightarrow$ associated with subsequent regulatory action
- Boilerplate deviation: unusual departure from peer templates
- Cross-modal: text strategy vs. quantitative factor exposures

### Multi-Modal Fusion

- NLP + quantitative returns: **+3–5%** $\mathrm{AUC}$ (Ahmed et al., 2024)
- Logic: text claims conservative equity L/S but returns load on leveraged distressed credit $\Rightarrow$ stronger signal

**Limitation**: filings are heavily boilerplate; low signal-to-noise ratio; fraudsters use compliance counsel to match expected patterns

**Source: Loughran & McDonald (2011); Araci (2019); Ahmed et al. (2024); paper Section 4.4**

# Graph Neural Networks for Fund Networks

### Architectures

- **GCN** (Kipf & Welling, 2017): spectral convolutions, neighborhood aggregation
- **GAT** (Velickovic et al., 2018): attention-weighted neighbor contributions
- **GraphSAGE** (Hamilton et al., 2017): inductive learning – score new funds without retraining (cold-start)
- **Temporal knowledge graphs**: time-stamped edges for evolving relationships

### Results from Adjacent Domains

- Wang et al. (2019): semi-supervised GNN on transaction networks – $\mathrm{AUC} = \mathbf{0.87}$
- Liu et al. (2021): "camouflage" problem – fraud detection even when immediate neighborhood appears benign

### Hedge Fund Application

- Service provider network: auditor, administrator, custodian, prime broker
- Small/unregistered auditors, lack of independent administrators $\Rightarrow$ elevated risk
- Dynamic signals: sudden auditor change, administrator linked to multiple sanctioned funds, manager "phoenix" pattern

### Strengths

- Captures relational info inaccessible to tabular methods
- "Guilt by association," network centrality, structural equivalence

### Limitations

- Graph construction requires entity resolution (largely unaddressed)
- Incomplete relationship data
- Computational cost scales with graph size

**Source: Wang et al. (2019); Liu et al. (2021); paper Section 4.5**

# Semi-Supervised and Self-Supervised Methods

**Label Propagation / Self-Training**

- Extend sparse labels through the data manifold
- Effective when $< 5\%$ of data carry labels (characteristic of hedge fund context)
- Self-training: iteratively label most confident predictions, retrain

**Contrastive Learning**

- Learn representations maximizing agreement between augmented views of same fund
- Minimize agreement between different funds
- Separate normal from anomalous without explicit fraud labels
- Downstream classification with very few labeled examples

**Self-Supervised Pre-Training**

- Objectives on unlabeled returns:
  - Masked return prediction
  - Temporal order prediction
  - Next-period forecasting
- Creates general-purpose fund behavior representations
- Fine-tune for fraud with minimal labels (pre-train $\rightarrow$ fine-tune paradigm)

**Transfer Learning**

- Adapt models from banking/insurance/accounting fraud
- Degree of cross-domain transferability $=$ open empirical question

**Limitation**: sensitive to distributional shifts; labeled examples may not represent full fraud population

**Source: Pang et al. (2021); paper Section 4.6**

# Synthetic Data and Data Augmentation

### SMOTE and Variants

- Most widely used: interpolation between existing positives in feature space
- Problem: interpolating between a Ponzi scheme and a valuation fraud generates implausible patterns
- Borderline-SMOTE, ADASYN: concentrate near decision boundary (partial fix)

### Conditional GANs

- Condition on fraud type or attributes
- Capture complex dependencies (returns $\times$ filings $\times$ operations)
- More realistic than interpolation

### VAEs

- Better-calibrated uncertainty estimates
- Advantageous when confidence of generated examples matters

### Validation Circularity

- Generator learns to resemble *known* fraud
- If known examples are not representative $\Rightarrow$ synthetic data perpetuates biases
- Ensuring realism *without* assuming we know what fraud looks like: fundamental challenge

### Synthetic Benchmarks

- Calibrated simulation (Fiore et al., 2019)
- Avoids confidentiality constraints
- Hedge-fund-specific benchmark = open priority (OP1)

**Source: Chawla et al. (2002); Fiore et al. (2019); paper Section 4.7**

# Method Families: Comparative Overview

method families: classical, tree-based, deep learning, NLP, GNN, semi-supervised, synthetic – showing performance ranges, data require

**Source: Paper Section 4**

# Critical Assessment: Reproducibility Crisis

**Proprietary Data**

- Majority of studies use **proprietary datasets**: licensed databases, internal regulatory records, bespoke compilations
- Results cannot be independently verified
- Performance metrics must be **taken on trust**
- Contrasts sharply with credit card fraud detection: public benchmarks enable rigorous comparison

**No Standard Benchmark**

- Each study assembles its own data, defines its own fraud labels
- Reports results on non-overlapping fund populations
- Cross-study comparison is effectively impossible
- An $\mathrm{AUC}$ of 0.79 in one study cannot be compared to 0.72 in another when:
  - Datasets differ
  - Label definitions differ
  - Feature sets differ
  - Evaluation protocols differ

**Source: Paper Section 4.8**

# Critical Assessment: Benchmark Gap and Domain Specificity

**Domain Specificity**

- Most impressive claims ($F_1 > 0.85$, $\mathrm{AUC} > 0.90$) originate from **adjacent domains**:
  - Credit card fraud
  - Payment fraud
  - Banking fraud
- These domains have abundant labeled data and well-characterized fraud
- Transferability to hedge funds is uncertain:
  - Sparse data
  - Heterogeneous fraud types
  - Sophisticated adversaries

**Class Imbalance Handling**

- Treatment varies enormously across studies
- Often inadequately reported:
  - SMOTE applied without impact evaluation
  - Ad hoc cost ratios for cost-sensitive learning
  - Some report only **accuracy** – meaningless under severe imbalance (classifying all as non-fraud $> 97\%$)
- No standardized protocols for handling *or* reporting class imbalance

**Studies reporting high performance on general financial fraud benchmarks likely overestimate effectiveness for hedge fund surveillance.**

**Source: Bolton & Hand (2002); Phua et al. (2010); paper Section 4.8**

# Critical Assessment: Evaluation Protocols and Publication Bias

**Temporal Evaluation Gap**

- Proper: train on period $t$, evaluate on period $t+1$ (temporal split)
- Many studies use random cross-validation
- Allows future information to leak into training set
- **Inflates reported performance**
- In a domain with concept drift, temporal evaluation is a *necessity*, not a refinement

**Publication Bias**

- Positive results preferentially published
- Studies reporting high detection rates more likely to reach peer-reviewed venues
- True performance landscape is **less optimistic** than published literature suggests
- Potential remedies:
  - Registered reports
  - Pre-registered analysis plans
  - Commitment to publish regardless of outcome
  - Not yet standard practice in this field

**Source: Paper Section 4.8**

# Critical Assessment: Overfitting Risk

**The Small-Sample Problem**

- Only $\sim$50–100 labeled fraud cases in the historical record
- Even moderate-complexity models risk **memorizing** idiosyncratic characteristics of specific schemes
- A model achieving high performance on 20 held-out fraud cases may simply have learned fingerprints of specific Ponzi schemes, valuation frauds, style misrepresentations in its training set
- Without capacity to detect novel fraud types

**Amplifying Factors**

- Common practice: single train-test split (rather than multiple independent evaluations)
- Heterogeneous positive class: Ponzi $\neq$ NAV manipulation $\neq$ insider trading
- Pooled fraud labels $\Rightarrow$ compromise decision boundary
- May fail to detect any individual fraud type with high sensitivity

**Mitigation Needed**

- Fraud-type-specific evaluation
- Multiple temporal splits
- Synthetic augmentation with validation
- Ensemble of type-specific detectors

**Source: Paper Section 4.8**

# Literature Gaps: Visual Summary

showing method families vs. critical challenges (reproducibility, benchmark, domain specificity, class imbalance, evaluation protocol, pub

**Source: Paper Section 4.8**

# Summary: Field at Early Stage of Scientific Maturity

1. **Classical methods** (Benford, serial correlation, omega-score) provide foundational features but each captures only one fraud dimension
2. **Tree-based ensembles** dominate: best performance ($F_1 \sim 0.88$), SHAP-compatible, robust to class imbalance
3. **Deep learning** offers theoretical appeal (temporal patterns, nonlinear representations) but faces severe data scarcity ($< 100$ positives) and overfitting risk
4. **NLP** (FinBERT 87%, SEC-BERT) enables text-quant cross-modal detection; multi-modal fusion adds **+3–5%** $\mathrm{AUC}$
5. **GNNs** capture relational "guilt by association" ($\mathrm{AUC}$ 0.87 in adjacent domains) but hedge fund application remains empirically underexplored
6. **Semi-supervised / self-supervised** methods directly address label scarcity through contrastive learning and pre-train/fine-tune paradigms
7. **Synthetic data** (GANs, VAEs) address class imbalance but face validation circularity
8. **Critical gaps:** no standard benchmark, reproducibility crisis, evaluation protocol inconsistencies, publication bias, and high overfitting risk with 50–100 cases

**Source: Paper Section 4**