Appendix B: Feature Engineering Details

AI-Based Detection of Hedge Fund Fraud

Comprehensive Feature Specifications

Feature Engineering Overview

Critical Role

Feature quality and informativeness directly determine model performance in hedge fund fraud detection systems.

Five Feature Categories:

1. **Statistical Features:** Return properties, risk metrics, distributional characteristics
2. **Benford Features:** Conformity to mathematical laws of digit distributions
3. **Textual Features:** Disclosure characteristics, readability, sentiment
4. **Network Features:** Relationship structures, centrality measures
5. **Temporal Features:** Time-varying patterns, regime changes, calendar effects

Typical Feature Space

50–200

features in practice

Dimensionality varies by:

- Data availability
- Model complexity
- Regulatory context
- Interpretability needs

Multi-faceted approach captures different aspects of fund behavior to identify fraud patterns

Featu

engineering is the cornerstone of effective fraud detection systems.

Return-Based Characteristics

Statistical properties of return time series reveal anomalies indicative of manipulation or fraud.

Feature	Formula/Description	Fraud Signal
First-order autocorrelation	$\rho_1 = \text{Corr}(r_t, r_{t-1})$	High ρ_1 indicates return smoothing
Sharpe ratio	$SR = \bar{r}/\sigma_r$	Abnormally high SR suggests fabricated returns
Maximum drawdown	$MDD = \max_t \left(\frac{\max_{s \le t} P_s - P_t}{\max_{s \le t} P_s} \right)$	Unusually low MDD indicates smoothing
Kurtosis	Excess kurtosis: $\text{Kurt}(r) - 3$	Extreme kurtosis signals manipulation
Skewness	Third moment of distribution	Asymmetry patterns reveal reporting bias
Discontinuity at zero	Kink in return distribution at zero (Bollen-Pool)	Avoidance of negative returns
Hurst exponent	H via rescaled range (R/S) analysis	Long-range dependence, non-randomness

Data Requirements: Minimum 24–36 months of monthly return data for reliable estimation

Getm

Mathematical Law Conformity

Naturally occurring financial data follows Benford's Law; deviations suggest manipulation or fabrication.

Benford's Law:

- Leading digit distribution is logarithmic
- Digit "1" appears 30% of the time
- Digit "9" appears 4.6% of the time
- Applies to many naturally occurring datasets

Expected Probability:

$$P(d) = \log_{10}\left(1 + \frac{1}{d}\right)$$

for first digit $d \in \{1, 2, \ldots, 9\}$

Feature Specifications:

Test	Formula
First-digit test	$\chi^2 = \sum_{d=1}^{9} \frac{(O_d - E_d)^2}{E_d}$
Second-digit test	Benford's law on second digit
Summation test	Cumulative conformity across positions

Applications:

- Return series
- NAV values
- Fee disclosures
- Transaction amounts

Amira

et al. (2015), Jorion et al. (2015). High χ^2 indicates fabricated data.

Disclosure Analysis

Textual characteristics of regulatory filings, annual letters, and offering documents reveal deception patterns.

Feature	Formula/Description	Fraud Signal
Fog index	$0.4[(w/s) + 100(c/w)]$ where w=words, s=sentences, c=complex words	High readability difficulty obscures risk
FinBERT sentiment	BERT-based financial sentiment score $\in [-1, 1]$	Overly positive sentiment masks problems
Boilerplate deviation	$1 - \text{CosSim}(\text{doc}, \text{template})$	Unusual language suggests concealment
Topic modeling	LDA or BERT-based topic distributions	Topic shifts indicate strategic framing
Disclosure length	Word count, section lengths	Excessive length obscures key information

Data Sources:

- Form ADV (U.S.), AIFMD disclosures (EU)
- Annual investor letters
- Offering memoranda and private placement documents

Preprocessing: Tokenization, stopword removal, domain-specific term handling

Brow

Relationship Structures

Network topology of fund-service-provider relationships reveals suspicious patterns and fraud contagion.

Network Structure:

- **Nodes:** Funds, managers, auditors, administrators, prime brokers
- **Edges:** Service relationships, ownership ties, board connections
- **Graph type:** Undirected or directed depending on relationship

Feature Specifications:

Feature	Formula
Degree centrality	$\sum_j A_{ij}$
Betweenness centrality	$\sum_{s \neq v \neq t} \frac{\sigma_{st}(v)}{\sigma_{st}}$
Related-party count	Number of disclosed relationships

Fraud Signals:

- **Low degree:** Isolated funds with few relationships
- **High clustering:** Tight-knit groups suggest collusion
- **Self-custody:** Fund controls its own assets
- **Related-party auditors:** Conflicts of interest
- **Offshore domiciles:** Opaque jurisdictions

Data Sources:

- Regulatory databases
- Commercial data vendors
- Manual extraction from disclosures

Challenge: Incomplete data due to selective disclosure

Zhan

et al. (2022), Cassar et al. (2015). Network features capture relational fraud patterns.

Time-Varying Patterns

Temporal dynamics, regime changes, and calendar effects reveal behavioral anomalies over time.

Feature	Description	Fraud Signal
HMM regime indicator	Binary state from Hidden Markov Model	Persistent "good" regime unrealistic
Change-point score	Bayesian change-point detection probability	Abrupt performance shifts suspicious
Calendar effects	Month-end, quarter-end, year-end patterns	Strategic timing of reported returns
Volatility clustering	GARCH-based conditional volatility	Abnormal clustering patterns
Trend strength	Moving average deviation	Unrealistic constant trends
Periodicity	Fourier-based cycle detection	Artificial periodic patterns

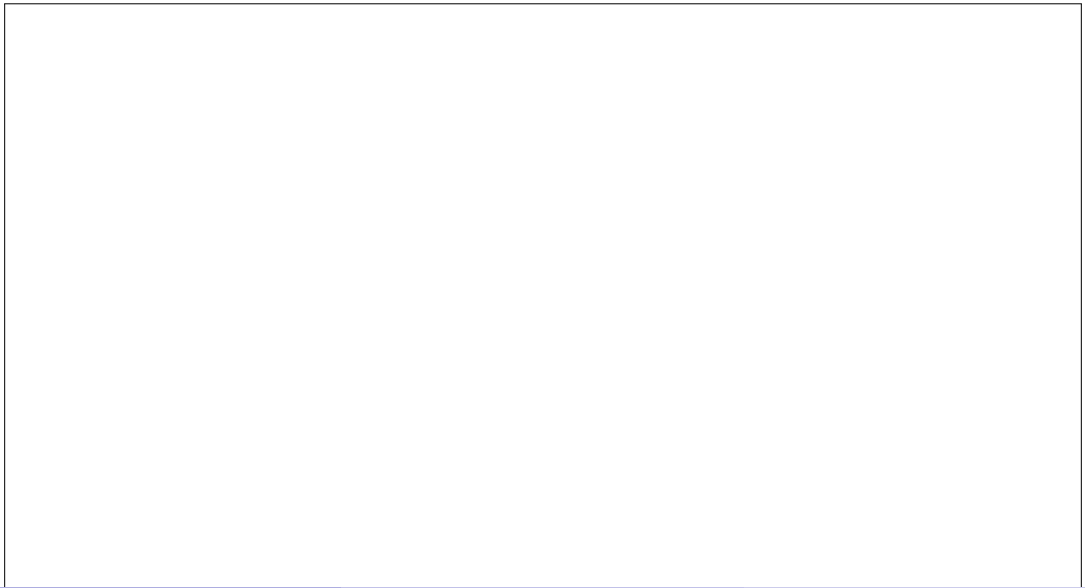
Rolling Window Computation:

- Enable real-time detection
- Maintain statistical power
- Typical window: 24–60 months

Temporal Modeling:

- LSTMs capture sequential dependencies
- HMMs identify regime switches
- Bayesian methods detect change-points

Comprehensive Feature Summary



Feature Selection and Dimensionality Reduction

Managing High-Dimensional Feature Spaces

With 50–200 features, selection techniques are essential for model performance, interpretability, and regulatory acceptance.

Feature Selection Methods:

1. **LASSO Regularization**
   - L1 penalty: $\lambda \sum_j |\beta_j|$
   - Automatic feature selection
   - Interpretable linear models
2. **Tree-Based Importance**
   - Random Forest, XGBoost
   - Gini importance, permutation importance
   - Non-linear feature interactions
3. **Mutual Information**
   - Information-theoretic criterion
   - Captures non-linear dependencies
   - Suitable for neural networks

Dimensionality Reduction:

- **PCA:** Less common due to loss of interpretability
- **Autoencoders:** Rare in regulatory applications
- **Preference:** Feature selection over reduction

Why Avoid Dimensionality Reduction?

1. **Interpretability:** Regulators require feature-level explanations
2. **Auditability:** Individual features must be traceable
3. **Legal requirements:** Explainability mandates (EU AI Act)
4. **Domain knowledge:** Selected features align with fraud theory

Best Practice: Combine domain expertise with data-driven selection for interpretable, high-performance models

Data Source Requirements by Category

Data Availability and Quality

Feature construction depends on access to high-quality, comprehensive data from multiple sources.

Category	Data Sources	Quality Challenges
Statistical	Monthly return series (24–36 months minimum)	Backfill bias, survivorship bias, missing data
Benford	Returns, NAV values, fee disclosures	Rounding effects, small sample sizes
Textual	Form ADV, AIFMD disclosures, annual letters, offering documents	Unstructured format, OCR errors, incomplete filings
Network	Regulatory databases, commercial vendors, manual extraction	Selective disclosure, evolving relationships, incomplete coverage
Temporal	Rolling return windows, market data	Requires long histories, computational complexity

Commercial Databases:

- Hedge Fund Research (HFR)
- Morningstar
- Preqin
- Bloomberg

Regulatory Sources:

- SEC EDGAR (Form ADV)
- EU AIFMD registers
- CFTC Form PF
- National regulators

Robu

fraud detection requires multi-source data integration and quality control.