An Integrated Portfolio Allocation System with Multi-Source Financial Intelligence and Constraint-Programming Feasibility Repair

Alan Salazar
Universidad Intercontinental de la Empresa
A Coruña, Spain
https://github.com/alanslzrr/financial-intelligence-lab

***Abstract*—Recent work on financial AI evaluates sentiment analysis, return prediction and portfolio optimization as separate components, with limited evidence on how these signals interact within a single constrained allocation. This paper asks whether a system that combines heterogeneous public signals can produce better portfolio decisions than passive strategies when it operates under formal rebalancing constraints.**

We build an integrated pipeline where FinBERT sentiment, per-ticker LSTM predictions and technical momentum signals are combined via cross-sectional z-scores and fed into a Genetic Algorithm optimizer. A CP-SAT solver then repairs each proposed allocation to satisfy six hard constraints (budget, position ceiling, position floor, sector cap, cardinality, turnover ceiling). We evaluate with an ablation walk-forward over 2022–2023 on 52 tickers, 30 random seeds per variant, and a nested bootstrap that captures both temporal and optimizer noise. All data is open (yfinance, FNSPID, FRED).

The full pipeline reaches the highest average Sharpe ratio across seeds (1.38). Under the conservative nested bootstrap no pairwise comparison reaches significance at 5% in the 501-day window. The strongest evidence is operational: the constraint repair layer cuts annual turnover from 131% to 35% with 100% feasibility across 96 monthly rebalances. Sentiment absorbs 60% of the signal weight; the LSTM takes 5%. A parallel CWT-scalogram exploration for visual volatility forecasting is a documented negative result: HAR-RV outperforms both ViT and CNN approaches.

***Index Terms*—portfolio optimization, constraint satisfaction, sentiment analysis, LSTM, genetic algorithm, financial intelligence, open data**

1. Introduction

1.1 Context and motivation

There is growing evidence that AI techniques such as sentiment classifiers, recurrent networks and metaheuristic optimizers can extract useful signals from financial data. Over the past two years the literature has expanded from isolated models to multi-step pipelines that summarize earnings calls, predict returns and draft investment reports. Components perform well in isolation. The step from a signal to a portfolio decision that respects real constraints is rarely addressed.

Four recent surveys and benchmarks converge on the same picture: the field is fragmented across applications rather than evaluated jointly within a single allocation [1]; multimodal integration often fails to add value over text alone [2]; multi-agent LLM pipelines still fit an assistant role with the human analyst keeping final control [3]; and most deep learning studies rely on proprietary data feeds that block reproduction [4].

From these readings we identified three gaps that frame the contribution of this work.

1.2 Problem statement

Gap 1 — Fragmented validation. Sentiment analysis, return prediction and portfolio optimization appear as separate research lines. There is limited evidence on what happens when these signals coexist within a single portfolio and compete for weight in the same allocation decision.

Gap 2 — Weak operational explicitness. Many proposals arrive at a recommended set of portfolio weights but rarely check whether those weights satisfy budget, cardinality, sector or turnover limits. The translation from "optimal" to "feasible" is often left implicit.

Gap 3 — Centrality of intermediate metrics. Measures like RMSE, F1 or R^2 describe a model internally but do not predict its effect on a portfolio-level objective such as Sharpe ratio or maximum drawdown.

Our research question: *Can a system that combines heterogeneous signals from public data produce better portfolio decisions than simple passive strategies when it operates under formal rebalancing constraints?*

The word *better* requires precision. We evaluate at the portfolio level using Sharpe ratio, drawdown and turnover, and test differences with a bootstrap that accounts for both temporal dependence and optimizer stochasticity. A parallel exploratory line asks whether CWT scalograms processed by a Vision Transformer or a compact CNN can

forecast realized volatility beyond the HAR-RV baseline [5]; we treat it as independent of the portfolio pipeline.

1.3 Contributions

1. **An integration pipeline with CP-SAT repair** that combines FinBERT sentiment, per-ticker LSTM predictions and technical momentum signals and produces formally feasible allocations. In the ablation, the constraint layer cuts annual turnover from 131% to 35% with 100% feasibility across 96 monthly rebalances.
2. **A nested-bootstrap ablation protocol** that resamples both temporal blocks and GA seeds (30 per variant). V4 achieves the highest seed-averaged Sharpe (1.38); no pairwise Sharpe comparison is significant at 5% in the 501-day window, so the strongest empirical claim is the operational one.
3. **A documented negative result on visual volatility forecasting.** Across four approaches (OCR, candlestick CNN, ViT dual-path, scalogram CNN) HAR-RV [5] remains the strongest model.
4. **Full reproducibility on open data.** yfinance, FNSPID [6], FRED; no proprietary feeds. Universe, windows, seeds and rebalancing convention are fixed in a versioned contract.

2. State of the Art

2.1 Fragmented validation of financial AI components

Sentiment analysis, return prediction and portfolio optimization have mature individual literatures. They are rarely evaluated as parts of the same allocation decision. Jadhav and Mirza [1] review 84 studies on LLMs in equity markets between 2022 and early 2025 and organize them by application (forecasting, sentiment, portfolio management, trading) and by technique (prompting, fine-tuning, multi-agent, reinforcement learning). Most papers work on a single application at a time. What is missing is evidence on how these signals interact when they compete for weight inside the same portfolio.

The FinCall-Surprise benchmark illustrates the difficulty of multimodal integration even when the data is synchronized. Shu et al. [2] assemble 2,688 earnings calls with aligned transcripts, audio recordings and slide decks, then evaluate 26 models across four families. Adding modalities often degrades performance: the benchmark pipelines concatenate representations trained in isolation rather than jointly, which limits their ability to extract complementary signals.

Giantsidi and Tarantola [4] survey 187 deep learning studies for financial forecasting published between 2020 and 2024, covering LSTMs, CNNs, Transformers and hybrid architectures across stock, index, forex, commodity and volatility tasks. They find that most work evaluates a single model on a single asset class, with limited attention to how the forecasting output connects to downstream portfolio decisions.

2.2 Weak operational explicitness in portfolio recommendations

A substantial number of papers propose portfolio weights or trading signals, but far fewer verify whether those proposals satisfy the constraints a real portfolio faces: budget balance, position limits, sector exposure caps, cardinality bounds, turnover ceilings. The implicit assumption is that a weight vector close to the optimal one can be rounded or adjusted later without much loss. DeMiguel, Garlappi and Uppal [7] show that this assumption does not hold robustly: across seven empirical datasets, none of 14 optimized strategies consistently beats the naive 1/N rule out of sample, in part because estimation error compounds through unconstrained optimization.

Jadhav and Mirza [1] note the gap between analytical outputs and actionable decisions as one of the open challenges for LLM-based financial systems. The Frontiers review calls for hybrid models that can handle real-world frictions, not just clean backtests. Satapathy et al. [3] attempt to move closer to decision support by generating structured investment reports from earnings calls through a multi-agent LLM system. Their evaluation measures report quality and finds that the system is best used as an assistant that highlights drivers, risks and scenarios, while human analysts keep final control of capital allocation. The authors themselves conclude that full automation of the investment decision is not yet reliable.

In the classical optimization literature, Markowitz [8] establishes mean-variance optimization as the theoretical foundation, but the original formulation does not include the integer and combinatorial constraints (cardinality limits, minimum lot sizes, turnover caps) that matter in practice. Constraint Programming solvers such as CP-SAT [9] can enforce these constraints exactly.

2.3 Centrality of intermediate metrics

Most evaluation in financial AI focuses on metrics that describe the model's internal accuracy. Common examples are RMSE for regressors, F1 for classifiers and R^2 for volatility forecasters. These metrics do not directly describe the effect of the model's output on a portfolio. A sentiment classifier with 90% F1 may contribute nothing to allocation quality if its errors concentrate on the tickers that carry the most weight. A price predictor with 51% directional accuracy may be valuable or useless depending on how its signal is weighted and constrained.

Fischer and Krauss [10] evaluate LSTM networks for financial market prediction on S&P 500 constituents from 1992 to 2015 and report daily returns of 0.46% and a Sharpe ratio of 5.8 before transaction costs. They also observe that from 2010 onwards the excess returns appear to have been arbitraged away: net returns hover near zero once realistic trading costs are applied. The gap between the model-level metric (daily return of the signal) and the portfolio-

level metric (net Sharpe after costs and constraints) is the problem we address.

Araci [11] introduces FinBERT and evaluates it on Financial PhraseBank classification accuracy. Subsequent work uses FinBERT scores as portfolio signals, but the connection between classification F1 and portfolio Information Coefficient is rarely made explicit.

2.4 Visual approaches to volatility forecasting

Our fourth objective explores whether visual representations of price dynamics carry predictive information about volatility. It is motivated by a separate but related line of work. Corsi [5] introduces the Heterogeneous Autoregressive model of Realized Volatility (HAR-RV), a linear model that regresses future volatility on lagged averages at daily, weekly and monthly horizons. Despite its simplicity, HAR-RV remains a strong baseline that many neural approaches struggle to beat.

Wooh and Cho [12] propose TF-ViTNet, a dual-path architecture that feeds Continuous Wavelet Transform (CWT) scalograms through a Vision Transformer while processing numerical features through an LSTM, fusing both streams for volatility prediction. Their results on NASDAQ and S&P 500 data show improvements over numerical-only models. This motivated our dual-path exploration.

However, our results diverge from theirs. On our 52-ticker universe with walk-forward quarterly evaluation, HAR-RV ($R^2 = 0.77$) outperforms both the LSTM-only model ($R^2 = 0.74$) and the dual-path ViT + LSTM ($R^2 = 0.73$). The visual path adds no information beyond what the numerical features already contain. We document this as a negative result. Section 5.3 discusses possible explanations, including smaller sample size, per-ticker heterogeneity and the choice of a ViT backbone pretrained on natural images.

3. Methodology

We follow a Design Science Research approach [13]. We design and iteratively evaluate a software artifact, the rebalancing system together with its supporting modules. The components respond to the three gaps identified in Section 2. The experimental protocol is fixed in a versioned contract that specifies the asset universe, time windows, seeds and rebalancing conventions before any results are generated.

3.1 System architecture

The **experimental layer** is a notebook pipeline that defines the controlled portfolio experiment: data ingestion, sentiment, prediction, optimization, visual volatility regression and the final ablation, plus a regime-detection classifier on the visual side (Section 3.5). The universe, calendar, seeds and evaluation metrics are locked before execution, and every stage produces versioned artifacts (parquet files, JSON logs, model checkpoints) that the downstream stages consume.

The **application layer** is a full-stack platform (FastAPI backend + Next.js frontend + PostgreSQL) that translates the useful parts into a usable product. The backend follows the same structure as the notebook pipeline, but it is not a byte-for-byte rerun of the experiment: it uses a lighter GA for latency, can use live APIs when keys are available, and falls back to cached notebook artifacts when live data is missing or rate-limited. The experimental results in Section 4 come from the notebook layer.

Data sources. All data is public and reproducible. Prices are daily OHLCV for 52 S&P 500 tickers from yfinance (January 2020 to March 2025, ~1,316 trading days per ticker). News are 158,617 articles from FNSPID [6], a dataset of 15.7 million financial news items published at KDD 2024 that covers S&P 500 companies from 1999 to 2023; we use the 2020–2023 subset, filtered by coverage (minimum 500 articles per ticker). Macro uses eight series from the Federal Reserve Economic Data (FRED) repository: Federal Funds Rate, 10-year and 2-year Treasury yields, VIX, USD Index, CPI, unemployment rate, 10Y–2Y spread — forward-filled to align with daily trading dates, not interpolated. The daily series (DFF, DGS10, DGS2, VIXCLS, DTWEXBGS, T10Y2Y) are indexed by observation date (also release date); CPI and UNRATE are indexed by reference month, so the feature matrix receives them ~10–15 days earlier than a real-time user would. Their combined weight in the 23-feature input is small (2/23), but the look-ahead is non-zero and we flag it as a threat to validity. Fundamentals are company-level snapshots from yfinance (sector, market cap) used for constraint definitions.

Time windows. Training: January 2020 to December 2021. Out-of-sample (OOS): January 2022 to December 2023 (501 trading days, 24 monthly rebalancing periods), covering a bear market (2022) and a recovery (2023). An anti-leakage rule enforces the separation: any training row whose 5-day label horizon falls within the OOS window is excluded.

3.2 Sentiment signal

Each headline is scored with FinBERT [11], a BERT fine-tuned on financial text, using a continuous score $s = p(\text{positive}) - p(\text{negative}) \in [-1, +1]$ that preserves intensity. Headlines are grouped by ticker and date, and for each group we record mean, median and standard deviation of s, article count, and the positive/negative fractions. Days without news have no sentiment row, so the sentiment signal contributes only when a headline exists.

The signal is validated at two levels. Classification reaches 97.2% F1 on Financial PhraseBank (circular, since FinBERT was trained on it) and 56.2% F1 on FiQA-2018 (unseen, domain mismatch lower bound). At the signal level we compute the Information Coefficient (IC) against

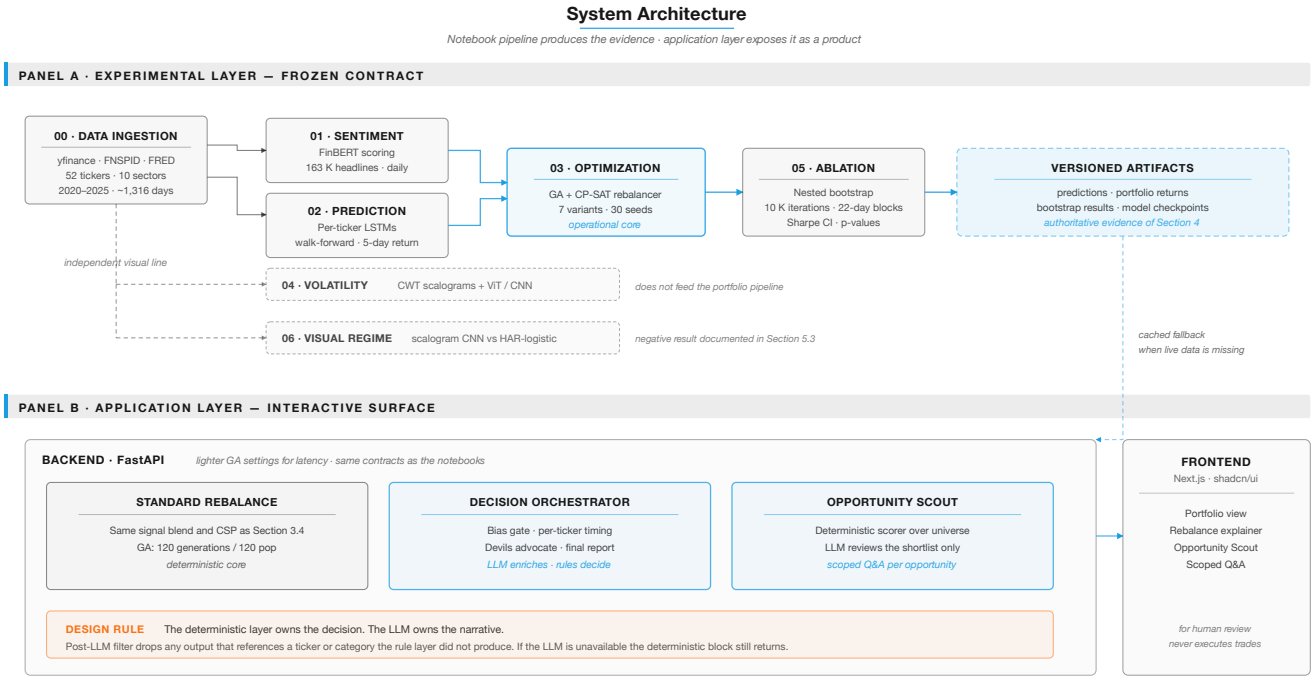

Fig. 1. System architecture. The notebook pipeline produces the evidence; the application layer surfaces it without rerunning the experiment. The visual volatility line is independent of the portfolio pipeline.

future returns under four methodologies: raw daily and rolling 5-day are not significant, extreme-only is significant at 5 days (IC = 0.025, $p = 0.023$), and cross-sectional IC is significant at 1 and 10 days. The signal carries marginal information in extreme readings and relative rankings, not in the raw daily time series.

3.3 Prediction signal

The target is the 5-day forward percentage return, $y_t = (p_{t+5} - p_t)/p_t$, which normalizes the scale across tickers and matches the input the optimizer expects. 23 features go into the model in three groups: 6 price-derived (returns at 1/5/20 days, 20-day rolling volatility, volume ratio, high-low range), 9 technical indicators (RSI(14), MACD line + signal + histogram, Bollinger %B and bandwidth, ATR(14), ATR%, OBV slope; all implemented from scratch to control warmup and NaN propagation), and the 8 FRED macro series.

We train one LSTM [14] per ticker, two layers, 64 hidden units, 60-day sliding window, ~56K parameters. A larger model would overfit on ~500–700 samples per walk-forward window, and a single pooled model would smooth across very different volatility profiles. Training uses an expanding window over 24 monthly steps on the OOS period; each month retrains on all available history with 20% held out for early stopping (patience 10). The scaler is refit at each window to prevent forward leakage through normalization statistics.

Baselines are naive (last observed 5-day return), ARIMA(5,1,0) and TimesFM 2.5 (200M-parameter zero-shot foundation model). We report RMSE, MAE and directional accuracy (DA), and test with a binomial on DA $> 50\%$, Diebold–Mariano on squared error and McNemar on directional accuracy.

3.4 Constrained rebalancing engine

The engine receives the three signals, combines them into a scoring function, proposes weight vectors through a Genetic Algorithm, and repairs each proposal to formal feasibility through a CP-SAT solver.

Signal combination. Each signal is normalized to cross-sectional z-scores winsorized at ± 3. The three are 20-day momentum (technical), 30-day average FinBERT score (sentiment) and the latest LSTM 5-day predicted return. The combined score for asset i is

$$\text{score}_i = w_{\text{tech}} \cdot z_{\text{tech},i} + w_{\text{sent}} \cdot z_{\text{sent},i} + w_{\text{pred}} \cdot z_{\text{pred},i}$$

The weights $(w_{\text{tech}}, w_{\text{sent}}, w_{\text{pred}})$ and a maximum turnover parameter were selected by a 570-combination grid search on the training window (2020–2021); the winning configuration is 0.35/0.60/0.05 with max_turnover = 0.30. Missing readings (mostly sentiment on ticker-days without news) are filled with zero after z-scoring, so the slot contributes its cross-sectional mean without redistributing its weight.

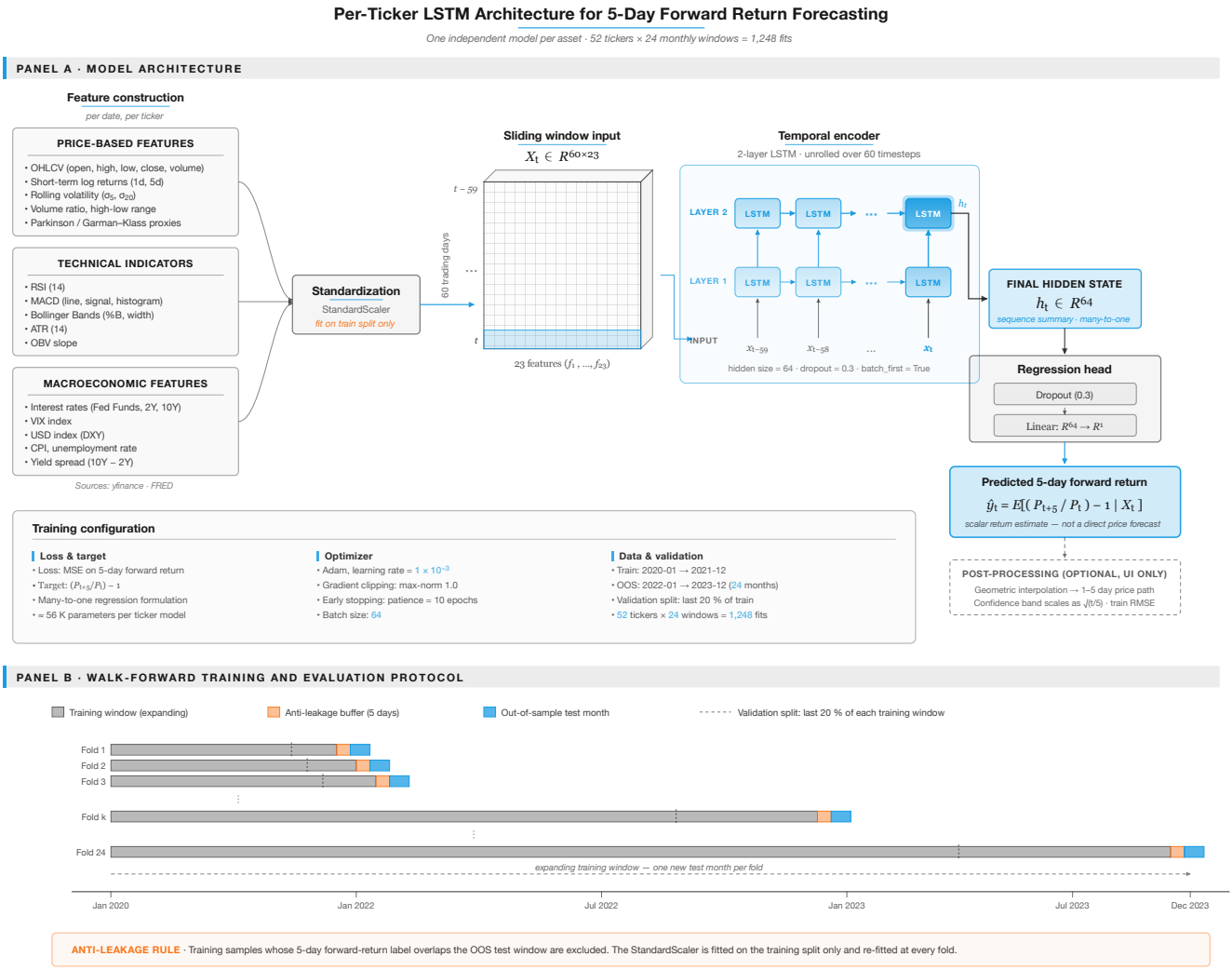

Fig. 2. Per-ticker LSTM architecture and walk-forward training protocol. The anti-leakage rule excludes any training row whose 5-day forward label overlaps the OOS window.

Optimizer. The covariance matrix for the fitness function is estimated with Ledoit-Wolf shrinkage [15] on a trailing 252-day window, with a diagonal fallback below 60 observations. The GA evolves a population of 200 weight vectors on the simplex. It uses tournament selection, BLX-α crossover, Gaussian mutation with a decaying σ and 5% elitism. Fitness is the Sharpe ratio from the combined signal scores and the shrunk covariance, adjusted for the number of active positions.

CP-SAT repair. We frame the repair step as a Constraint Satisfaction Problem (CSP) and solve it with the CP-SAT solver from Google OR-Tools [9]. The solver receives the GA's proposed weight vector and finds the nearest feasible allocation in L_1 distance subject to six hard constraints: (1) budget — weights sum to 100%; (2) maximum weight per asset — no single asset above 25%; (3) minimum weight if held — any held asset above 3%; (4) sector cap — no sector above 40%; (5) cardinality — at most 20 assets simultaneously; (6) turnover ceiling — total absolute weight change from the previous period at most max_turnover. The solver works in integer arithmetic and finds exact solutions. We keep the GA and the CSP as two stages on purpose: the GA explores the space freely and the CSP repairs the final candidate once. Embedding the hard constraints inside the GA would render most offspring infeasible.

Rebalancing convention. Monthly rebalancing on the first business day. Weights are computed using data available up to date t and take effect from $t + 1$. The one-step-ahead application is the structural guard against lookahead in the backtest. A minimum trade threshold $|\Delta w| < 0.005$ rounds residual rebalances to zero.

Ablation variants. Seven variants are constructed to isolate the contribution of each signal family and the constraint layer, as summarized below.

B0 and B1 are deterministic. V1–V5 are stochastic (GA) and run with 30 random seeds each. V4 is the full pipeline; V5 isolates the CSP's effect by using the same signals without repair.

3.5 Visual volatility exploration

An independent experimental line that does not feed the portfolio optimizer. It tests whether CWT scalograms carry predictive information about future Parkinson volatility [16] beyond numerical features. Each 60-day volatility window becomes a 2D Morlet scalogram (24×60), fed in two ways: a dual-path model following Wooh and Cho [12] (frozen ViT-B/16 [17] pretrained on ImageNet + LSTM on HAR features, ~162K trainable parameters, regression head); and a compact CNN (~40K parameters) that classifies high-volatility regimes against the trailing 252-day 75th percentile. Baselines are naive persistence, HAR-RV [5] for regression, and HAR-logistic for classification. Protocol: eight quarterly walk-forward windows over 2022–2023, thresholds tuned on the last training quarter.

3.6 Statistical evaluation

The primary test is a circular block bootstrap with 22-day blocks ($\approx$ one trading month, chosen to preserve monthly autocorrelation) and 10,000 iterations, extended to a nested design: each iteration samples a seed from the 30-seed matrix for each stochastic variant, then resamples temporal blocks from that seed's return series. This captures market noise and GA solver noise in a single CI. For deterministic baselines (B0, B1) only temporal resampling applies. Pairwise tests reuse the same bootstrap indices within each iteration so compared variants face the same synthetic market. We also report an inter-seed CI (range of Sharpe across 30 seeds) that measures GA convergence consistency; the nested CI is the primary test.

3.7 Application layer

The pipeline is exposed as an interactive surface (FastAPI + Next.js) that reuses the validated defaults with a lighter GA for latency and serves cached notebook artifacts when live data is unavailable. Two LLM-augmented surfaces sit on top: a decision orchestrator that narrates a rebalance with rule-based bias flags, timing guardrails and a deterministic traffic light, and an opportunity screener that ranks buy candidates from six deterministic signal components and only then asks the LLM for prose. In both, the deterministic layer owns the decision and the LLM is restricted to the rule-produced set. These surfaces are not evaluated in Section 4 and do not affect any number reported there.

4. Experiments and Results

Results follow the protocol in Section 3. The universe (52 tickers, 10 sectors), windows (training 2020–2021; OOS 2022–2023, 501 days, 24 monthly rebalances) and anti-leakage rule follow Section 3.1. The universe is coverage-conditioned by the FNSPID filter and is not representative of the full tradable space. Each experiment below reports the component result (4.1 sentiment, 4.2 prediction); the integrated ablation and CSP comparison are in 4.3; statistical significance in 4.4; the visual volatility exploration in 4.5.

4.1 Experiment 1: Sentiment signal

We score 163,099 FinBERT headlines (after deduplication and ticker aliasing) into 42,044 daily observations, median 793 days/ticker, range 587 (MSI) to 9,854 (TSLA). Distribution: 65% neutral, 19% positive, 17% negative, mean +0.040.

Raw daily IC is flat. Only extreme readings and cross-sectional rankings reach significance, consistent with a signal that contributes inside a blend rather than alone. Under Bonferroni ($\alpha = 0.05$, 12 configs) only the 10-day cross-sectional result survives. We read the IC table as exploratory; the portfolio-level grid search (Section 4.3) is the confirmatory check, and it assigns sentiment 60% of the weight.

4.2 Experiment 2: Price prediction

We evaluate the per-ticker LSTM forecaster against three reference models (naive, ARIMA, TimesFM) on 5-day forward returns. The walk-forward protocol produces 1,248 fits (52 tickers $\times$ 24 monthly windows), 26,052 predictions.

ARIMA and TimesFM fall below chance; the LSTM is the only model exceeding 50%. The binomial rejects 50% but assumes an independence that overlapping 5-day labels violate. McNemar ($p = 0.56$) shows no directional edge over naive; Diebold–Mariano confirms lower squared error. The LSTM gains magnitude calibration, not direction, and the 50.81% aggregate is modest because the signal is uneven across the universe (30 of 52 tickers clear 50% DA). A 200M zero-shot foundation model does not beat a 56K domain-specific LSTM here. The grid search assigns the LSTM 5% of the weight, in line with that contribution.

4.3 Experiment 3: Portfolio optimization and ablation

We evaluated 570 combinations of $(w_{\text{tech}}, w_{\text{sent}}, w_{\text{pred}}, \text{max\_turnover})$ on the training window (2020–2021) using seed 42. The winning configuration is $w_{\text{tech}} = 0.35$, $w_{\text{sent}} = 0.60$, $w_{\text{pred}} = 0.05$, max_turnover = 0.30. The grid search selects the blend that maximizes portfolio Sharpe, not the one that reproduces per-signal IC: a signal with weak univariate IC can still help through cross-sectional ranking, which is why sentiment takes 60% of the weight despite the raw-daily IC being near zero.

V1–V5 values are means over 30 GA seeds; B0 and B1 are deterministic. Sentiment drives the step V1 $\rightarrow$ V2 (ΔSharpe = +0.385); prediction alone adds nothing (V1 $\rightarrow$

Variant	Signals	Constraint repair	Seeds
B0	Equal weight (1/N monthly)	—	Deterministic
B1	Buy and hold	—	Deterministic
V1	Technical	GA + CSP	30
V2	Technical + sentiment	GA + CSP	30
V3	Technical + prediction	GA + CSP	30
V4	Technical + sentiment + prediction	GA + CSP	30
V5	Technical + sentiment + prediction	GA only (no CSP)	30

Benchmark	Weighted F1	Note
Financial PhraseBank	97.2%	Circular — FinBERT was trained on this dataset
FiQA-2018	56.2%	Unseen; social media text, domain mismatch provides a lower bound

Methodology	Horizon	IC	p-value	Significant?
Raw daily	1d	≈ 0	> 0.10	No
Rolling 5-day	5d	≈ 0	> 0.10	No
Extreme-only	5d	0.025	0.023	Yes
Cross-sectional	1d	0.013	0.029	Yes
Cross-sectional	10d	0.017	0.007	Yes

Model	RMSE	MAE	Dir. Accuracy	N
ARIMA(5,1,0)	0.0696	0.0463	34.35%	26,052
TimesFM 2.5 (200M params, zero-shot)	0.0710	0.0473	48.75%	26,052
LSTM pipeline (64h/60d)	**0.0826**	**0.0568**	**50.81%**	**26,052**
LSTM tuned (128h/180d)	0.0908	0.0635	49.64%	25,885
Naive (last 5d return)	0.0990	0.0665	50.55%	26,052

Test	Statistic	p-value	Interpretation
Binomial (LSTM DA $> 50\%$)	—	0.0047	LSTM beats chance
McNemar (LSTM vs Naive direction)	$\chi^2 = 0.33$	0.5633	Not significantly different
Diebold-Mariano (LSTM vs Naive, squared error)	DM $= -15.81$	< 0.001	LSTM has significantly lower squared error

Variant	Sharpe	Return	Volatility	Max DD	Turnover
B0 (equal weight)	0.610	13.0%	25.2%	−25.9%	8.4%
B1 (buy and hold)	0.582	11.2%	22.5%	−22.2%	—
V1 (tech + CSP)	0.911	21.7%	23.4%	−20.9%	55.1%
V2 (tech + sent + CSP)	1.296	26.0%	19.9%	−16.8%	54.7%
V3 (tech + pred + CSP)	0.927	18.1%	19.4%	−17.2%	55.0%
V4 (full pipeline)	**1.384**	**25.7%**	**18.6%**	**−15.6%**	**34.8%**
V5 (full, no CSP)	1.126	20.4%	18.1%	−16.0%	131.1%

V3, +0.016). V4's +0.088 over V2 is modest, so sentiment dominates the selected blend.

The V4 vs V5 comparison isolates the CP-SAT effect:

The CSP cuts turnover by a factor of 3.8× while the Sharpe point estimate increases by 0.26. Every V1–V4 portfolio satisfies the five structural constraints on all 96 rebalances, and the turnover ceiling on the 92 rebalances where it binds (the four opening moves from cash sit outside $\|\Delta w\|_1$). The Sharpe gap is not significant under the nested bootstrap; the turnover and feasibility result is the more defensible claim. The tighter turnover also caps portfolio velocity during stress: V4 posts the shallowest drawdown of the group.

4.4 Experiment 4: Statistical significance

Circular block bootstrap with 22-day blocks, 10,000 iterations, nested design (each iteration samples both a seed and a temporal block sequence).

No comparison reaches 5%; V4 vs B1 is the closest ($p = 0.10$). V4's nested bootstrap CI is $[-0.03,\ 2.96]$; the lower bound touches zero, so we cannot reject Sharpe = 0 over 501 days despite the 1.38 point estimate. With ~23 effective independent draws (22-day blocks over 501 days) the test is underpowered by construction.

The inter-seed CIs of V4 ($[1.29,\ 1.47]$) and V5 ($[1.11,\ 1.14]$) do not overlap: the GA consistently finds higher-Sharpe feasible portfolios for V4 across all 30 seeds. This is a statement about the optimizer, not the market. The operational claims — V4's point-estimate ordering and the 100% feasibility / 3.8× turnover cut — are deterministic and reproduce across seeds.

4.5 Experiment 5: Visual volatility

Regression (eight quarterly walk-forward windows, 13,004 samples, target 5-day forward Parkinson volatility) and a binary high-volatility-regime classifier against the trailing 252-day 75th percentile:

HAR-RV and HAR-logistic win on every metric. The dual-path model does not improve over the LSTM-only branch; the visual path adds no information the numerical features do not already contain. CNN AUROC ranges 0.502–0.638 across quarters and decision thresholds span 0.20–0.79, consistent with an unstable model. Two earlier visual attempts were abandoned without a formal benchmark: OCR on EDGAR filings (no stable public corpus) and a candlestick-chart CNN for price direction (collapsed to the majority class). The visual path never outperforms a numerical baseline; possible reasons are discussed in Section 5.

5. Discussion

5.1 How each gap was addressed

Fragmented validation. The ablation tests what happens when signals usually studied in isolation coexist in the same portfolio. Sentiment is weak alone (raw daily IC near zero), yet inside the blend it becomes dominant: 60% of the weight in the optimal configuration and the largest Sharpe step from V1 to V2 (+0.385). Under the nested bootstrap this step is not significant, so we read it as a useful ranking signal in this window, not proof of a stable market advantage.

Weak operational explicitness. The CP-SAT layer addresses this gap directly. Every V1–V4 portfolio satisfies all six constraints across 96 rebalances with 100% feasibility, and turnover drops from 131% to 35% vs. V5, which under realistic transaction costs separates an executable strategy from one that is not. The Sharpe gap V4 vs V5 (+0.26) is not significant ($p = 0.58$); the operational improvement is the feasibility and turnover control itself.

Centrality of intermediate metrics. The mismatch between model-level and portfolio-level metrics is visible in the LSTM: 50.81% directional accuracy is above chance under the binomial ($p = 0.0047$), yet V3 (technical + prediction) is indistinguishable from V1 (technical alone) at portfolio level, ΔSharpe = +0.016, $p = 0.96$. The binomial looks promising alone; the ablation shows the signal does not survive at portfolio level.

5.2 Integration results

V4 has the best point estimate of Sharpe (1.384) and the shallowest drawdown (−15.6%) among all variants. Under the nested bootstrap no pairwise comparison reaches 5%; the closest is V4 vs buy-and-hold at $p = 0.10$.

The interpretation depends on which claim is being made.

- **Point estimates and ordering.** V4 > V2 > V5 > V3 ≈ V1 > B0 > B1. Adding sentiment helps, adding prediction alone does not, and adding constraint repair reduces turnover without sacrificing return.
- **Operational properties.** 100% feasibility, turnover reduced by 3.8×, and inter-seed reproducibility (V4's $[1.29, 1.47]$ does not overlap with V5's $[1.11, 1.14]$). These results are deterministic and do not depend on window length.

The strongest empirical claim sits in the operational layer.

5.3 The visual volatility result

Four approaches to visual information from financial data were tried, and all four lost to numerical baselines:

1. OCR on scanned documents. Discontinued for lack of a stable public corpus on EDGAR.
2. Candlestick chart classification. Models collapsed to the majority class.
3. ViT dual-path volatility regression. HAR-RV $R^2 = 0.77$ vs. dual-path $R^2 = 0.73$.
4. CNN regime classification. HAR-logistic AUROC = 0.72 vs. CNN AUROC = 0.61.

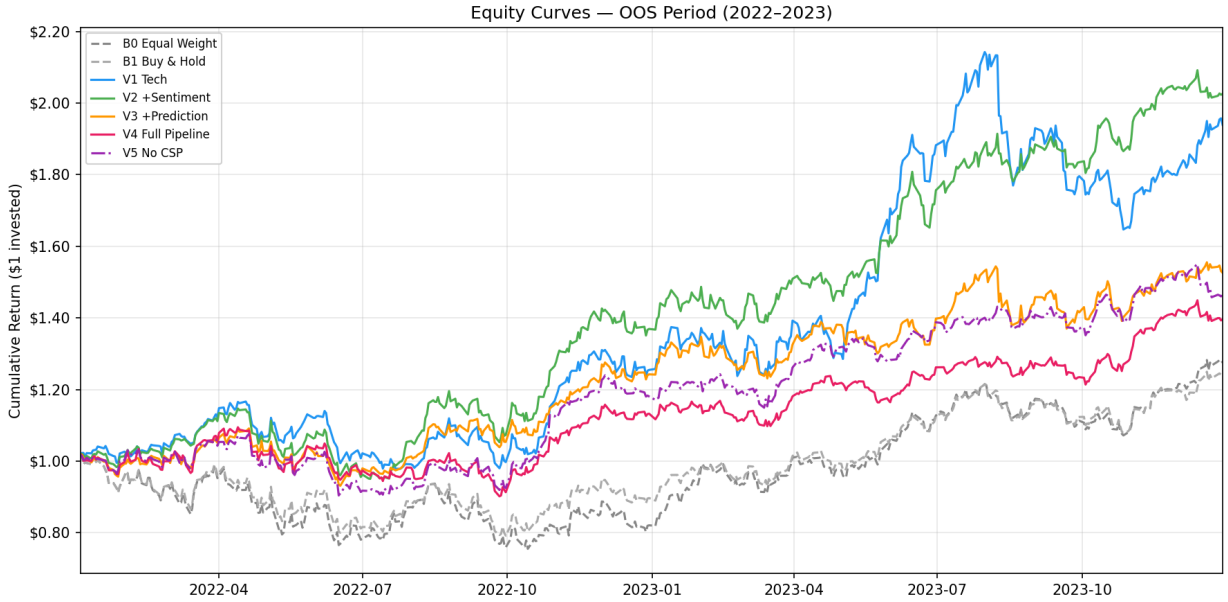

Fig. 3. Cumulative equity curves for the seven variants over the OOS period (2022–2023).

Metric	V4 (with CSP)	V5 (without CSP)
Sharpe (seed-mean)	1.384	1.126
Annual turnover	34.8%	131.1%
Max drawdown	−15.6%	−16.0%
Feasibility (96 rebalances)	100%	N/A (unconstrained)

Comparison	Δ Sharpe	p-value	Significant at 5%?
V4 vs B1 (full pipeline vs buy-and-hold)	+0.80	0.10	No
V4 vs B0 (full pipeline vs equal weight)	+0.77	0.17	No
V4 vs V5 (CSP effect)	+0.27	0.58	No
V2 vs V1 (sentiment effect)	+0.40	0.55	No
V3 vs V1 (prediction effect)	+0.03	0.96	No
V4 vs V1 (combined signals effect)	+0.49	0.48	No

Model	R^2	AUROC
Naive persistence	0.714	—
HAR-RV / HAR-logistic [5]	**0.774**	**0.717**
LSTM-only (numerical)	0.742	—
Dual-path (ViT + LSTM)	0.731	—
Scalogram CNN	—	0.607

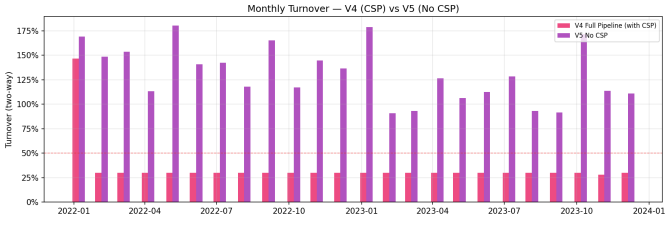

Fig. 4. Monthly turnover — V4 (with CSP) vs V5 (without CSP). The gap holds every month; the first V4 bar (2022-01) is the opening move from cash.

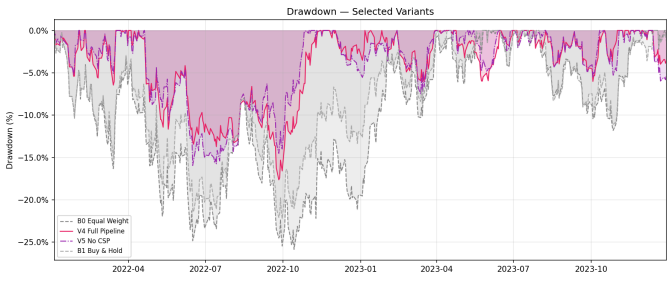

Fig. 5. Drawdown curves for B0, B1, V4 and V5 over the OOS period.

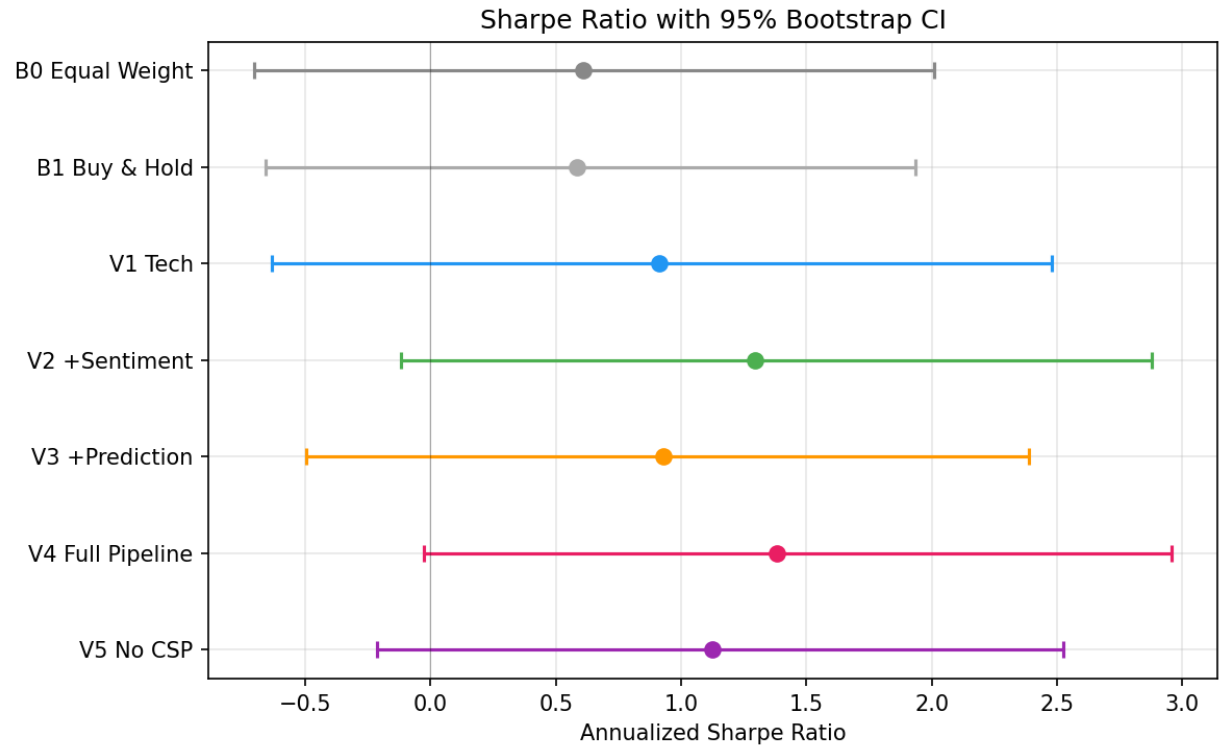

Fig. 6. Sharpe ratio confidence intervals (nested bootstrap, 95%) for all seven variants. CIs overlap across the group — the 501-day OOS window does not resolve the pairwise differences.

Likely explanations:

- **Redundant encoding.** The CWT scalogram transforms the same volatility series that the numerical branch already sees. If it adds no structure beyond lagged averages, the vision path processes a noisier version of the same information.
- **Domain mismatch.** The ViT backbone was pretrained on ImageNet. Financial scalograms lack the edges and textures the ViT learned to detect. A backbone pretrained on spectrograms might fare differently.
- **Sample size.** ~13,000 samples across 52 tickers and 8 quarterly windows is modest for a vision model.
- **HAR-RV is a strong baseline.** Corsi's model [5] exploits the multi-scale persistence of volatility. Our visual representations did not add information beyond that.

This result contrasts with Wooh and Cho [12], who report improvements with a similar architecture on S&P 500 and NASDAQ index-level data. The difference may reflect our per-ticker evaluation (52 diverse tickers vs. a single index), our smaller sample size, or differences in the scalogram construction.

5.4 Limitations

1. **OOS window length.** 501 trading days produce wide CIs under block bootstrap and limit statistical power. A longer window or a rolling multi-period design would help.
2. **No live trading.** The system produces recommendations but has not been tested with a live execution layer. Slippage, partial fills and latency are not captured in the backtest.
3. **Prediction horizon.** The LSTM targets 5-day returns. Other horizons may produce different signal-weight configurations; we did not search over them.
4. **Fixed signal weights.** Weights are selected on 2020–2021 and applied to 2022–2023. A shift in the relative informativeness of sentiment, prediction and momentum would degrade the fixed blend.

5.5 Practical implications

Following Satapathy et al. [3], the system produces recommendations for human review rather than executing trades. The constraint repair layer ensures each recommendation is already feasible, so the analyst no longer needs the manual adjustment step that typically follows an optimization run. The LLM-augmented surfaces in Section 3.7 apply the same rule we apply to the optimizer: the deterministic layer owns the decision, the LLM owns the interpretation.

5.6 Threats to validity

Internal validity. Three classes of leakage are easy to introduce in this kind of pipeline: temporal leakage through forward-shifted labels that overlap OOS, lookahead when

freshly computed weights are applied to same-day returns, and selection bias when hyperparameters are chosen on the OOS window itself. Section 3 describes the structural guards that prevent each one.

External validity. Results are specific to 52 U.S.-centric tickers over 2022–2023. Generalization to other markets, time periods or asset classes cannot be assumed, and the sentiment signal inherits FNSPID's coverage.

Construct validity. The Sharpe ratio penalizes upside and downside volatility symmetrically. The Sortino ratio (reported in the ablation output but not used as primary test) penalizes only downside volatility and may better reflect investor preferences.

6. Conclusions

6.1 Answer to the research question

The research question asked whether a system combining heterogeneous public signals can produce better portfolio decisions than passive strategies under formal rebalancing constraints. The answer is partially affirmative. The full pipeline produces the best point estimate of risk-adjusted return and the shallowest drawdown, with an internally consistent ordering across variants, and the CP-SAT repair layer delivers a measurable operational improvement. The 24-month OOS window is not long enough to reject the null at 5% for any pairwise Sharpe comparison. The operational contribution therefore rests on deterministic evidence; the performance contribution stays as a best point estimate with inconclusive significance.

The visual volatility line closes as a negative result: across four attempts — OCR, candlestick CNN, ViT dual-path regression and scalogram CNN classification — no visual representation beat the HAR baselines. The integration pipeline shows, by contrast, that signals weak in isolation (sentiment IC near zero, LSTM DA 50.81%) can contribute inside a constrained blend when the constraint layer removes the turnover cost that would otherwise absorb the gains.

6.2 Future work

1. **Longer evaluation window.** Extending OOS beyond 24 months or using a rolling multi-period design across market regimes would raise the statistical power of the bootstrap.
2. **Adaptive signal weighting.** The current grid search is static. A mechanism that adjusts weights in response to changing signal quality (e.g. reinforcement learning) could improve robustness.
3. **Domain-adapted language models.** FinBERT was not designed for the cross-sectional ranking task that drives its portfolio contribution. A model fine-tuned on financial news with a ranking objective might produce a stronger signal.
4. **Visual volatility with domain-specific backbones.** A ViT pretrained on financial spectrograms, or self-supervised methods on the scalogram corpus itself, could change the outcome of the visual experiments.

7. Code and data

All data sources (yfinance, FNSPID, FRED) are public. The experimental protocol — asset universe, time windows, seeds, rebalancing convention, signal weights, nested bootstrap parameters — is fixed in Section 3 and in the versioned contract that accompanies the code repository.

Repository. Code, data-fetch scripts and notebooks are hosted on GitHub,[1] with commits tagged at each experimental milestone. Python dependencies are pinned in `notebooks/requirements.txt` for the experimental pipeline and in `backend/requirements.txt` for the application layer. The notebooks expect Python 3.11; the LSTM training stage benefits from a CUDA-capable GPU but does not require one.

Reproduction. Running notebooks 00–05 in order — data ingestion, sentiment, prediction, optimization, visual volatility regression, and the ablation — regenerates every artifact consumed by Section 4. Each notebook writes versioned parquet / JSON outputs that the next stage reads, so partial re-runs are supported. The three stochastic components (GA optimizer, nested bootstrap, LSTM initialization) are seeded; the 30-seed list for the ablation is in the repository's experimental contract. Wall-clock time on a single consumer GPU is dominated by the LSTM stage (1,248 per-ticker fits) and the nested bootstrap (10,000 iterations over 7 variants); the remaining stages run in minutes.

References

[1] A. Jadhav and V. Mirza, "Large language models in equity markets: Applications, techniques, and insights," *Frontiers in Artificial Intelligence*, vol. 8, p. 1608365, 2025. [Online]. Available: https://www.frontiersin.org/journals/artificial-intelligence/articles/10.3389/frai.2025.1608365/full

[2] D. Shu, Y. Liu, H. Zhang, and M. Du, "FinCall-Surprise: A large scale multi-modal benchmark for earning surprise prediction," *arXiv preprint arXiv:2510.03965*, 2025. [Online]. Available: https://arxiv.org/abs/2510.03965

[3] R. Satapathy, R. Liew, J. Chattorj, E. Cambria, and R. S. M. Goh, "From earnings calls to investment reports: Evaluating role-based multi-agent LLM systems," in *Proceedings of the 10th Workshop on Financial Technology and Natural Language Processing*. Association for Computational

[1] https://github.com/alanslzrr/financial-intelligence-lab

Linguistics, 2025, pp. 258–267. [Online]. Available: https://aclanthology.org/2025.finnlp-2.19/

[4] S. Giantsidi and C. Tarantola, "Deep learning for financial forecasting: A review of recent trends," *International Review of Economics & Finance*, vol. 104, p. 104719, 2025. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1059056025008822

[5] F. Corsi, "A simple approximate long-memory model of realized volatility," *Journal of Financial Econometrics*, vol. 7, no. 2, pp. 174–196, 2009.

[6] Z. Dong, X. Fan, and Z. Peng, "FNSPID: A comprehensive financial news dataset in time series," in *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2024, pp. 4918–4927.

[7] V. DeMiguel, L. Garlappi, and R. Uppal, "Optimal versus naive diversification: How inefficient is the 1/N portfolio strategy?" *The Review of Financial Studies*, vol. 22, no. 5, pp. 1915–1953, 2009.

[8] H. Markowitz, "Portfolio selection," *The Journal of Finance*, vol. 7, no. 1, pp. 77–91, 1952.

[9] L. Perron, F. Didier, and S. Gay, "The CP-SAT-LP solver (invited talk)," in *Proceedings of the 29th International Conference on Principles and Practice of Constraint Programming (CP 2023)*, ser. LIPIcs, vol. 280, 2023, pp. 3:1–3:2.

[10] T. Fischer and C. Krauss, "Deep learning with long short-term memory networks for financial market predictions," *European Journal of Operational Research*, vol. 270, no. 2, pp. 654–669, 2018.

[11] D. Araci, "FinBERT: Financial sentiment analysis with pre-trained language models," *arXiv preprint arXiv:1908.10063*, 2019.

[12] M. Wooh and P. Cho, "Integrating vision transformer and time–frequency analysis for stock volatility prediction," *Mathematics*, vol. 13, no. 23, p. 3787, 2025.

[13] A. R. Hevner, S. T. March, J. Park, and S. Ram, "Design science in information systems research," *MIS Quarterly*, vol. 28, no. 1, pp. 75–105, 2004.

[14] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[15] O. Ledoit and M. Wolf, "A well-conditioned estimator for large-dimensional covariance matrices," *Journal of Multivariate Analysis*, vol. 88, no. 2, pp. 365–411, 2004.

[16] M. Parkinson, "The extreme value method for estimating the variance of the rate of return," *The Journal of Business*, vol. 53, no. 1, pp. 61–65, 1980.

[17] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," in *International Conference on Learning Representations (ICLR)*, 2021. [Online]. Available: https://openreview.net/forum?id=YicbFdNTTy