POLITECNICO
MILANO 1863

SCUOLA DI INGEGNERIA INDUSTRIALE
E DELL'INFORMAZIONE

Executive Summary of the Thesis

Does Multi-Source Data Beat Price Alone? A Controlled Ablation for Next-Day Stock Direction Prediction

Laurea Magistrale in Computer Science and Engineering - Ingegneria Informatica

Author: Mouadh Ltifi

Advisor: Prof. Fabrizio Pittorino

Co-advisors: Francesco Puoti, Prof. Molnár Bálint

Academic year: 2025-2026

1. Introduction

The multi-source financial-prediction literature is full of positive results. Methods that combine price with news, social-media sentiment, macroeconomic indicators, or inter-stock graph structure, fused through tensor products, gated cross-attention, or graph neural networks, regularly report gains over price-only baselines [3, 5]. To a practitioner, the message is that more data, suitably fused, buys skill.

The empirical-asset-pricing literature, looking at large cross-sections of US equities, paints a different picture: most published return predictors fail to replicate, and the survivors decay sharply once published [1]. This sets a low prior: a new predictability claim on liquid US equities should be doubted until it survives disciplined evaluation. The two records are usually read apart, and they are not flatly contradictory, because they test different things. What has rarely been done is to test the multi-source claim against that prior in one controlled ablation, with a single universe, a tuned price-only baseline, and corrected statistics.

This article runs that test, on the hardest case for the hypothesis: does integrating heterogeneous data sources improve *next-day directional prediction* (does a stock close up or down tomorrow?) for 55 large-cap US equities over 2015–2023, holding everything but the data source under test fixed? Three research questions structure it: whether multi-source integration beats price (RQ1), each source's marginal contribution (RQ2), and the conditions under which integration helps or hurts (RQ3). The answer matters in practice, since it decides whether scarce engineering effort should go into acquiring and fusing more data, or into the estimator and its evaluation.

The study makes two contributions:

- **A robust null.** Across two prediction targets, four text encoders, three fusion strategies, and two architectural families spanning a tenfold capacity range, no heterogeneous source significantly improves over price alone; the single non-null marginal effect is a *degradation* from backward-looking news.
- **A reproducibility audit.** Two prominent reported positives whose code is public are reproduced and shown to rest on evaluation choices, not data: MSGCA's headline tracks model selection performed on the test set, and CAMEF's holds only under a non-temporal split, one that trains on data from

the test period's future.

Together they give concrete reasons why the field's reported gains and the null reported here can coexist, and reorder a practical priority: estimator and evaluation rigour ahead of data sources.

2. Why a null is the expected outcome

Three features of the problem make a careful null the expected outcome, and fix its value.

A low prior on liquid US equities. The base rate for return predictability on this universe is low and independently established: most published anomalies fail to replicate under a common protocol (Table 1). A fresh positive on a heavily-studied US universe is therefore weak evidence until it clears disciplined evaluation, and the expected result of a careful test is a null.

Large-caps are the hardest case, so a null is an upper bound. The universe is deliberately adversarial to the hypothesis: large-cap US equities have the deepest liquidity, heaviest analyst coverage, and fastest absorption of public information into price, so if an alternative source carries exploitable residual signal anywhere, it should carry the *least* here. A null on this universe is therefore an upper bound on the field's positive claims: a positive here would have generalised to less efficient markets far more readily than the reverse.

Prior systems report gains with no shared baseline. The positive record is real but not mutually comparable (Table 1): each landmark system fixes a different market, horizon, and label rule, and often an untuned baseline, so a reported "gain over baseline" is rarely a gain over a *tuned price-only* model. The signal side reinforces the expectation: backward-looking text carries little forward information at the daily horizon, because by the time news enters a dataset the price has already moved on its content. What has been missing is the common denominator that the apparatus of Section 3 supplies.

3. Experimental design

The study is a falsification-oriented controlled ablation: a fixed prediction task, a single price-only baseline, and a grid of source additions, each judged against a statistical bar fixed in advance.

Universe and target. The universe is 55 large-cap US equities — the five largest by market capitalisation in each of the eleven GICS sectors, held fixed across 2015–2023 as a sector-balanced panel so no industry dominates the pooled estimate. The target is the sign of the next-day log return, with a symmetric 0.5% dead-zone that drops near-flat days so each retained label reflects a decisive move. The primary metric is the Matthews correlation coefficient (MCC), bounded in $[-1, +1]$ with 0 at chance and, unlike plain accuracy, not inflated by class imbalance. Because daily direction is genuinely hard, even a real edge yields only a small positive MCC, a few thousandths of a point, and the statistics below are built to resolve an effect that small. A secondary target, next-day realised volatility scored by R^2, tests whether the null is specific to direction.

Temporal folds. Evaluation uses five expanding-window folds (F0–F4, Figure 1): each trains on all data up to a cut-off and tests on the disjoint window that follows, so no future information enters training, and each training set strictly contains the previous one's (roughly 23,000 to 54,000 stock-days). The five test windows span distinct market regimes, which matters for a null: a source that helped in only one regime would surface in its fold, and every configuration gets five such chances. Early stopping uses the chronologically last 20% of each fold's training window as validation; the test set is read once.

The nine-configuration ablation. The baseline P is price plus ten standard technical indicators, a deliberately *strong* baseline (momentum, mean-reversion, volatility regime), so a candidate must beat enriched price, not raw price. Four sources are added to it (`N` news, FNSPID via frozen FinBERT and Qwen3; `M` macro, six FRED series plus an FOMC-window flag; `S` social, StockTwits; `G` an inter-stock graph, GICS co-membership plus rolling-correlation edges via a graph attention network). The nine configurations are a purposive subset of the sixteen possible, chosen so each research question is answered by a specific contrast:

Table 1: The evidence landscape. Three representative multi-source systems report gains (top); two independent results set the prior under which a careful null is expected (bottom). Because each fixes a different market, horizon, label rule, or baseline, the reported gains are not mutually comparable — the gap the common-denominator test here closes.

Study	What it reports	Relation to this study
MSGCA [5]	price + news + inter-stock graph (gated cross-attention); MCC ≈ 0.11 on BigData22	Audited in §5.
CMTF [3]	price + news + financials + macro (tensor fusion); direction gains on FTSE 100	A multi-source claim on a small universe; we test it at scale.
CAMEF [4]	price + macro-event text (causal); event-window gains around FOMC	Reproduced and re-evaluated in §5.
Koval et al. [2]	forward-looking text carries signal; backward-looking text does not	Explains why backward-looking news degrades our prediction.
Hou et al. [1]	452 published anomalies; 65–82% fail to replicate	Sets the low prior under which a careful null is the expected result.

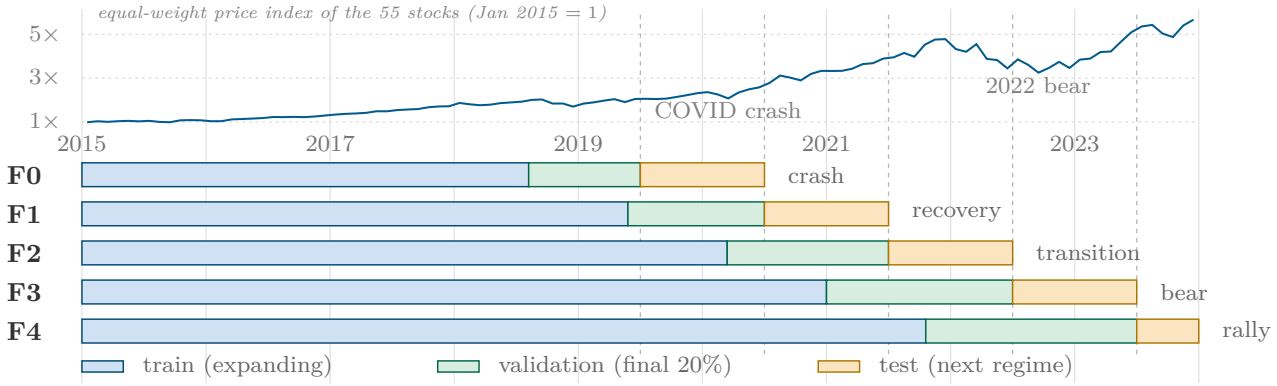

Figure 1: The five expanding-window folds on the 2015–2023 calendar, to scale, over the equal-weight price index of the 55-stock universe (January 2015 = 1). Each training span (blue) extends to its cut-off and contains the previous fold's; its final 20% (green) is held out for validation, and each test span (orange) is the disjoint window that follows. The fold-to-fold swing dwarfs the per-source effects under test — why every comparison is paired within a (fold, seed) cell.

- `P` — the baseline (price + technical indicators);
- `P+N`, `P+M`, `P+S`, `P+G` — each single source over price (the RQ2 add-one-in tests);
- `P+N+M`, `P+M+S`, `P+M+G` — each source anchored over macro;
- `P+N+M+S` — the joint ceiling.

Every non-price source is lagged to be observable at prediction time (macro releases shifted to their publication schedule), so no future value leaks backward.

Two architectures. A null is only as strong as the architectures that fail to break it, so two structurally distinct families run under the same grid. The *classifier-shaped* family is the dominant published shape: light per-modality encoders, a fusion module (concatenation, gated cross-attention, or multi-head self-attention), and a thin two-layer head (9,251–88,035 parameters); two variants differ only in how price enters (a one-day feedforward vector or a 20-day LSTM sequence). The *forecaster-shaped* family inverts that shape, placing ten times the parameters (195,000–278,000, 95–99% of them) in a Temporal Fusion Transformer body that derives both direction and volatility. The two bracket the tenfold capacity range on which the "too weak to find the signal" objection turns.

Statistical protocol. A model's MCC swings far more from one market period to another than from adding a source, so each configuration is compared to the baseline in matched pairs on the (fold, seed) cell, and it is the within-pair difference that is measured; because both

models met the same market, shared luck cancels. Every contrast reports the paired effect size Cohen's d (the gap in units of its run-to-run scatter; ≈0.2 small, 0.5 medium, 0.8 large), the Bonferroni-corrected p-value (which tightens the threshold for testing $k = 8$ configurations at once), and the sample size n; a finding counts only when all three agree, against a bar fixed before the forecaster family ran ($p_{\text{bonf}} < 0.05$, $|d| \geq 0.3$, fold consistency, an extension-seed retest). At the primary $n = 30$ resolution the design has 80% power to detect $d \geq 0.70$, which is why the null is re-checked across four robustness axes: a small effect might hide from one underpowered test but not from all. The grid totals 2,468 valid experiments across two independent codebases over a shared data pipeline, each run reproducible bit-for-bit from a fixed seed, every fold-dependent transform fit on training data only.

4. Results

No source beats price (RQ1, RQ2). The baseline is itself only weakly predictive: price plus technical indicators scores a pooled mean MCC of 0.0095, marginally above chance ($t = 2.061$, $p = 0.048$): the small edge efficient-market reasoning predicts, and the bar a useful source must clear. None clears it (Figure 2): every paired delta lies within ±0.010 of zero with a confidence interval straddling it, six of the eight are negative, and the contrast nearest significance ($p_{\text{bonf}} = 0.14$, P+N+M) makes prediction *worse*. The forecaster with ten times the capacity reproduces the null exactly. For RQ2 the picture is equally flat: the two largest single-source effects (macro $d = -0.25$, news $d = -0.19$) are both negative, and removing news or social from P+N+M+S changes nothing. Within the design's resolution, ranking the sources by usefulness is ranking noise.

The null survives every robustness axis. A single null is weak; one that survives every reasonable attempt to break it is strong. The null withstands four independent "but you didn't try . . . " objections, each a row of Table 2: a different prediction target (next-day volatility), a different text encoder (four spanning a 5.5-fold size range), a different fusion strategy (the three are statistically equivalent), and an order of magnitude more capacity. None clears

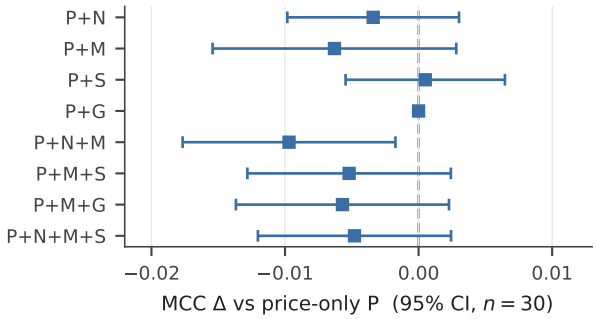

Figure 2: Paired MCC Δ of each configuration against the price-only baseline P ($n = 30$ per row, Bonferroni $k = 8$); points are deltas with 95% CIs. Every interval straddles the dashed zero line, the cluster tilting faintly negative: no configuration beats price, and the contrast nearest significance ($p_{\text{bonf}} = 0.14$, P+N+M) degrades prediction.

Bonferroni in either direction (the nearest miss is $p_{\text{bonf}} = 0.69$ on volatility, the closest positive $d = +0.593$ at $p_{\text{bonf}} = 0.339$ from the tenfold-larger forecaster), so the bottleneck is the information in the sources, not the encoder, the fusion, or the capacity. The one reliable architectural effect is orthogonal to the data question: pooled, the LSTM edges the feedforward variant ($d = 0.28$), extracting marginally more from the *price* sequence where both stay near chance.

Two bounded effects inside the null. Two effects survive (Table 2), neither a multi-source gain over price. First, *backward-looking news degrades* prediction over macro: the P+N+M-vs-P+M contrast is negative in four of five folds, largest in the two highest-news-flow regimes (the 2022 bear, the 2023 rally). The mechanism is timing: daily news sentiment correlates with same-day returns ($r = 0.143$) but barely with next-day ($r = 0.005$), so by the time an article enters the dataset the price has moved on its content. This is the forward-versus-backward distinction of Koval et al. [2], and the reason a better encoder does not help. Second, *social sentiment helps an LSTM*, but through the price model: giving the social stream its own temporal sequence yields no improvement, so the gain is the recurrent estimator extracting more from the *price* sequence, not new information in the posts. Neither social configuration beats price-only, and neither effect transfers to the forecaster: both concern the estimator and

Table 2: Every result in one place. "null" = no significant effect after Bonferroni; "equivalent" = no within-family difference; "bounded" = a real but conditional effect, not a multi-source gain over price. For the two bounded rows, d and p are fold-level ($n = 5$); the cell-level figures treating (seed, fusion, architecture) as independent ($d = -0.35$, $p_{\text{bonf}} = 0.010$ news; $d = +0.44$, $p = 0.006$ social) overstate the certainty.

Finding	RQ	d	p_{bonf}	Verdict
Any config beats price (classifier family)	RQ1	≤ 0.46	≥ 0.14	null
Any config beats price (forecaster / TFT)	RQ1	—	—	null
Marginal value of each source (add-one-in)	RQ2	≤ 0.25	—	null
Fusion strategy (concat / gated / MHSA)	RQ1	—	0.522	equivalent
Volatility target instead of direction	RQ1	≤ 0.32	≥ 0.69	null
Text-encoder choice (4 encoders)	RQ1	≤ 0.21	> 0.25	null
Capacity: TFT (10× params) vs classifier	RQ1	+0.59	0.339	null (closest)
Backward-looking news over price+macro	RQ3	−0.80	0.15	bounded (degrades, 4/5 folds)
Social sentiment under an LSTM	RQ3	+1.0	0.083	bounded (estimator, not source)

the data's timing, not the marginal value of an added source.

5. Reconciling with the positive literature: a reproducibility audit

If no source helps here, why does the literature so consistently report that one does? Because some of these systems release their code, the gap can be examined directly: one can ask whether, once the evaluation is free of leakage (any way test-period information slips into training or model selection), the reported positive still holds. We do this for two prominent systems, reporting the outcome as a reproduction result rather than a verdict on the system in its own setting. This is the study's second contribution. It identifies two distinct ways a reported gain can fail to survive disciplined evaluation.

MSGCA: the gain tracks test-set model selection. MSGCA [5] fuses price, news, and an inter-stock graph by gated cross-attention (one source's features re-weight another's) and reports a headline MCC of 0.1112 on BigData22. Its released training loop draws a random train/test split with *no validation set*, scores the test set every epoch, and keeps the maximum, so model selection is performed on the test set. Running the released code unmodified reproduces the headline (best-on-test ≈ 0.097 at the 300–400-epoch budget); changing *only* the selection rule to a held-out validation set collapses the same model and data to chance (Figure 3). Part of that drop is the smaller training set the validation arm must carve; the clean, like-for-like effect, with data and training length fixed, is that test-set selection inflates MCC by +0.04 to +0.07, stable across budgets. The reported positive does not survive leakage-free selection. This is one reproduction, not a survey, but it isolates a mechanism that recurs wherever a protocol leaves model selection unspecified — and leakage-free selection is the discipline this study applies to itself.

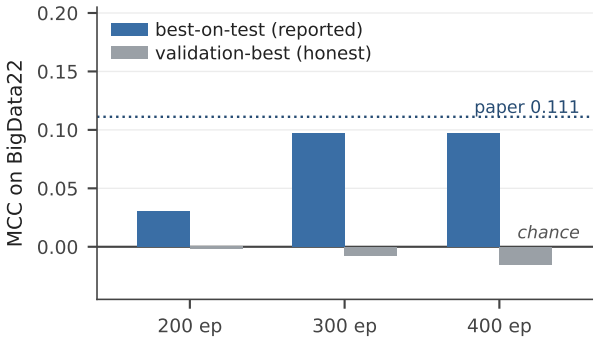

Figure 3: MSGCA on BigData22: the reported headline does not survive a leakage-free protocol. The released code (best-on-test, no validation set) reproduces the headline (≈ 0.097, near the paper's 0.1112); choosing the epoch on held-out validation collapses the same model and data to chance. With data fixed, selection alone accounts for +0.04 to +0.07 MCC.

CAMEF: a genuine headline that holds only under a non-temporal split. CAMEF [4] forecasts US assets around macroe-

conomic announcements and also releases its code; its checkpoint selection is sound (it minimises validation MSE), so the concern is the split. The audit runs in three steps. *The headline is real*: the released weights give MSE 0.000431 through the authors' own harness, matching the reported 0.000489. *But the documented recipe does not produce it*: a from-scratch retrain at the default five epochs reaches only 0.00249 ($\approx 5.8\times$ worse), so the released weights come from a longer, undocumented budget. *And the result depends on a non-temporal split*: a flag off by default and absent from every documented command governs whether the split respects time; by default ≈99.7% of training events post-date the earliest test event, so the model trains on the test period's future. Flipping it to chronological makes the authors' own recipe diverge, with test MSE blowing up to about 15. A collapse this severe makes it implausible that more training would rescue a chronological split: the headline rests on a non-temporal evaluation, of doubtful validity for a time-series forecast.

These two evaluations differ from this study's along axes common across the literature: one fixed split, a single best run, no multiple-testing correction, selection not always insulated from the test set. None means the systems are wrong in their own terms; it means the null reported here should be read as honest, not weak.

6. Conclusions

On next-day direction prediction of 55 liquid US large-caps, integrating heterogeneous data sources does not beat a price-only baseline. The answer to all three research questions is the same: no source carries positive marginal value (the one non-null effect is a *degradation* from backward-looking news), and the only conditions under which a source moves the result are adverse or estimator-bound, so no condition makes integration help. Because the universe is the hardest case, the null is a tight upper bound, not a weak average. Economically the values are immaterial: the best directional accuracy anywhere is about 52%, the price-only baseline is the single highest-Sharpe (best risk-adjusted return) configuration, and a delta at the resolution floor maps to roughly 50.15% accuracy, far below any plausible transaction-cost threshold.

The factors that *do* move the reported number are methodological, not informational: the fusion architecture is irrelevant, tenfold more capacity changes nothing, and the two reproduced positives owe their headlines to evaluation, not data (test-set model selection in one, a non-temporal split in the other). On this problem, getting the estimator and its evaluation right matters more than adding data sources, and a study that does not separate the two effects can mistake an evaluation artefact for an information gain. For a practitioner, spend effort on the estimator and honest out-of-sample evaluation before acquiring more data; for a reviewer, treat a multi-source "gain" as unverified until shown over a tuned price-only baseline with the test set read once, never used for model selection.

The boundaries of the null are deliberate, and point to the most informative future work, which shares one variable: *when* information reaches the model relative to when price has discovered it. Forward-looking text (earnings calls, guidance, analyst revisions) and the intraday horizon both increase the not-yet-absorbed component, and could finally yield a positive where the daily horizon does not. Until then, the default for next-day prediction on liquid equities is the one this study could not beat: price, evaluated honestly.

Acknowledgements

The author thanks Prof. Fabrizio Pittorino, Francesco Puoti, and Prof. Molnár Bálint for their supervision and guidance.

References

[1] Kewei Hou, Chen Xue, and Lu Zhang. Replicating anomalies. *Review of Financial Studies*, 33(5):2019–2133, 2020.

[2] Ross Koval, Nicholas Andrews, and Xifeng Yan. Financial forecasting from textual and tabular time series. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 8289–8300, 2024.

[3] Yunhua Pei, John Cartlidge, Anandadeep Mandal, Daniel Gold, Enrique Marcilio, and Riccardo Mazzon. Cross-modal temporal fusion for financial market forecasting. *arXiv preprint arXiv:2504.13522*, 2025.

[4] Yang Zhang, Wenbo Yang, Jun Wang, Qiang Ma, and Jie Xiong. CAMEF: Causal-augmented multi-modality event-driven financial forecasting by integrating time series patterns and salient macroeconomic announcements. In *Proc. ACM SIGKDD Conf. on Knowledge Discovery and Data Mining (KDD '25)*, 2025.

[5] Chang Zong and Hang Zhou. Stock movement prediction with multimodal stable fusion via gated cross-attention mechanism. *arXiv preprint arXiv:2406.06594*, 2024.