

Perturbation-Specific Reliability of Intrinsic and Post-Hoc Explainable Machine Learning in Volatility Forecasting

Haikal Djauhari

haikal@infiga.net

Universitas Ahmad Dahlan

Anton Yudhana

Universitas Ahmad Dahlan

- Sunardi

Universitas Ahmad Dahlan

Research Article

Keywords: Intrinsic interpretability, post-hoc explainability, volatility forecasting, explanation stability

Posted Date: October 26th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11158626/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

Perturbation-Specific Reliability of Intrinsic and Post-Hoc Explainable Machine Learning in Volatility Forecasting

Haikal Djauhari^{1*}, Anton Yudhana², Sunardi³

¹Doctoral Program in Informatics, Universitas Ahmad Dahlan, Yogyakarta, Indonesia 55191

²Doctoral Program in Informatics, Universitas Ahmad Dahlan, Yogyakarta, Indonesia 55191

³Doctoral Program in Informatics, Universitas Ahmad Dahlan, Yogyakarta, Indonesia 55191

*haikal@infiga.net

Abstract

The reliability of intrinsic and post-hoc explanations cannot be assessed independently of the perturbation under which they are evaluated. Using purged walk-forward validation, this study forecasts five-day realized volatility for 20 equities across three markets from 2016 to 2025 and evaluates five model families paired with intrinsic or post-hoc explanation methods. We admit only models that outperform step-5 persistence out of sample to the main interpretability analysis. We assess explanation reliability under three perturbations: retraining with different random seeds, temporal shift, and input jitter, using matched-stock paired inference. Integrated Gradients applied to a Long Short-Term Memory (LSTM) network and to TabNet differ by 0.52 in temporal rank agreement, showing that explanation stability depends on the model–representation–explainer configuration. When the same trained TabNet network is explained using its intrinsic attention mask and post-hoc Integrated Gradients, the two methods are similar under temporal shift but diverge under input perturbation and retraining. The intrinsic mask is more robust to input noise but less reproducible across retraining. Ablation-based faithfulness performs no better than random feature ordering, while fixed-depth top-k agreement is biased by feature-space size. The Heterogeneous Autoregressive model of Realized Volatility (HAR-RV) achieves the best median forecasting accuracy and is statistically indistinguishable from the spline-based Generalized Additive Model (GAM). Overall, explanation reliability is perturbation-specific and should be evaluated empirically for each model–representation–explainer configuration.

Keywords: Intrinsic interpretability; post-hoc explainability; volatility forecasting; explanation stability.

1 Introduction

Financial time-series volatility is often considered easier to forecast than returns. However, this remains difficult because of volatility clustering, heavy tails in return distributions, non-stationarity across market regimes, and asymmetric responses to positive and negative shocks [1–3]. Although machine learning (ML) models are able to deal more effectively than traditional methods such as GARCH or HAR with nonlinear relationships, high-dimensional interactions, and dynamics that depend on the market regime, they are not always better [2, 4]. The need for large amounts of data hinders their performance, increases the risk of overfitting, raises the computational costs linked to the repeated re-estimation using rolling windows, and results in inconsistent improvements out of sample when compared to simpler benchmark models [1, 5]. The accuracy of the forecasts also relies heavily on how the features are constructed, particularly in terms of choosing the predictors, representing the different market regimes, and regarding the level of redundancy or collinearity among the variables [6–10]. Another limitation is interpretability: whereas the coefficients of classical linear models have fairly

clear statistical and economic interpretations, many machine learning models function as opaque "black boxes", which hampers their acceptance by regulators, their use in risk management, and the ability to identify model failures during periods of extreme market conditions [11–13]. This issue has led to the adoption of explainable artificial intelligence (XAI), which includes both intrinsic and post hoc methods. Intrinsic methods build interpretability into the model itself, examples being generalized additive models (GAMs) [14], Explainable Boosting Machines (EBMs) [15], and TabNet [16]. On the other hand, post-hoc methods provide explanations after training and can be used with more complex architectures. Common techniques include SHAP and TreeSHAP [17, 18], counterfactual explanations [19], and Integrated Gradients [20]. Together, these methods offer complementary ways to increase transparency while maintaining the predictive flexibility of current machine learning models.

In fact, the vast majority of the explainability tools employed in both industry and academic research are of a post-hoc nature: SHAP, LIME, and counterfactual explanations can be used with any model that has already been trained without necessitating any modifications to the model, and it is for this reason that they are so attractive since they do

not place any extra demands on the model selection process [21–25]. Rudin [26] argues that this convenience is also the source of the problem: a post-hoc explainer approximates a model which it has not taken part in constructing, and such an approximation might misrepresent the very reasoning that it claims to describe — a risk which she regards as unacceptable in the case of important decisions, such as those involving finance. Laugel et al. [27] provided direct empirical evidence of this problem in counterfactual explanations, showing that many classifier predictions lie in regions of the input space that are not continuously connected to real training data of the same class. The authors term these regions “unjustified” because explanations generated there may reflect artifacts of the model’s decision boundary—such as isolated pockets from an unstable ensemble or extrapolated regions from an overly flexible classifier—rather than patterns genuinely supported by the data. Even though their study is focused on counterfactual explanations rather than the feature-attribution methods that are the subject of this paper, the fundamental concern is directly applicable: the output of a post-hoc explainer can be affected by the idiosyncrasies of the trained model rather than by real structures in the data, and there is nothing in the post-hoc method itself that guarantees this will not occur.

Yet the debate remains unresolved, even though Rudin [26] and Laugel et al. [27] have drawn a distinction. Wiratsin et al. [28] and Haque et al. [29] claim that neither intrinsic nor post-hoc explanations have consistently shown faithfulness, that is, that an explanation truly matches the model’s real reasoning. Regarding intrinsic methods, genuinely interpretable architectures remain rare in practice, and many models that are supposed to be interpretable still include black-box components. For instance, attention weights can be obtained from representations already transformed by uninterpreted encoders and thus may not indicate the true importance of the input. More generally, assertions about intrinsic interpretability are often accepted without adequate validation, attention being a clear case of potentially unfaithful explanations as discussed by Jacovi & Goldberg [30]. Post-hoc methods also suffer from similar issues. Feature-attribution techniques can differ greatly, and their faithfulness usually depends on the particular task and model rather than being generally applicable. Recent theoretical research has even indicated that it may be impossible to guarantee generally faithful post-hoc importance measures, as shown by Bilodeau et al. [31]. Likewise, similar limitations have been found in relation to neuron-level visualizations and probing methods, since explanations that appear convincing do not necessarily correspond to the model’s actual

internal mechanisms. Some literature therefore questions whether the distinction between intrinsic and post-hoc approaches should remain the main framework for explainability. Rather than assuming one paradigm is inherently trustworthy, this study treats faithfulness, stability, and consistency as empirical properties to assess for each model–representation–explainer configuration.

The idea put forward by Rudin [26] that intrinsically interpretable models should be given preference over post hoc explanations of black-box models remains disputed. She maintains that the supposed trade-off between accuracy and interpretability is generally misleading, noting that interpretable models can achieve predictive performance competitive with other models. Yet in many practical situations, post hoc explanations are still needed, since high-performing models may already be in use, may be proprietary, or may be too expensive to replace with intrinsically interpretable ones. In these circumstances, post hoc methods might be the only reasonable way to achieve transparency. However, their reliability is not guaranteed. Intrinsic models share a similar drawback: their explanations are trustworthy only if the public model structure truly reflects the reasoning behind the predictions. Rather than taking this property for granted, the present study assesses it empirically.

The existing XAI benchmarks generally do not tackle the question of whether explanations are intrinsic or post-hoc and instead evaluate different kinds of explainers (for instance, SHAP with XGBoost rather than attention in a transformer) — in this way they determine which model is more interpretable rather than which explanation method is more reliable, since the differences between the base models interfere with any differences that are observed. To properly assess Rudin’s recommendation, we need both an intrinsic explanation and a post-hoc explanation for the same trained model, so any discrepancy between the two can be attributed to the explanation method rather than to one model being easier to interpret than the other. To our knowledge, no previous study has carried out such a comparison in the context of a financial forecasting task, nor has it examined it under conditions that are realistic for that domain — specifically, non-stationary markets, walk-forward evaluation, and a skill gate that makes sure the model being explained actually has predictive ability from the beginning.

The paper addresses that gap directly. The paper compares intrinsic and post-hoc explanations on the same model under controlled conditions: TabNet produces an intrinsic explanation via its attention masks, and the same trained TabNet model is then given a post-hoc explanation using Integrated

Gradients. Therefore, any difference between the two methods must come from the explanation technique, not the model. This comparison is part of a more comprehensive cross-model benchmark that includes five different model families — XGBoost, an LSTM, TabNet, a spline-based GAM, and an EBM — each being paired with the explainer appropriate to its architecture and each being evaluated in the context of 5-day-ahead realized volatility forecasting for 20 equities from the US, European, and Japanese markets, using a leakage-safe walk-forward procedure. Rather than relying on a single interpretability criterion, we assess each model–representation–explainer configuration using three complementary metrics: faithfulness, stability (with respect to input noise and over time), and consistency across independent explainers for the same model.

Accordingly, this study makes three methodological contributions. First, it isolates explanation effects through a same-model comparison in which intrinsic TabNet masks and post-hoc Integrated Gradients explain the same fitted network. Second, it decomposes explanation stability into input-noise, temporal, and re-estimation perturbations under leakage-safe walk-forward evaluation. Third, it evaluates explanation quality with randomized faithfulness controls and size-corrected agreement measures so that apparent reliability is not inferred from metric artifacts.

2 Related Work

2.1 Explainable Artificial Intelligence in Financial Time Series

Arsenault [32] examined the use of XAI over a five-year period in financial time series forecasting and developed a comprehensive taxonomy, distinguishing explainability from interpretability—two notions often conflated but used differently in practice. This distinction underpins the current study’s comparison between intrinsic interpretability, which applies to models designed to be interpretable (e.g., GAMs and EBMs), and post-hoc explainability, which refers to techniques applied after training to explain black-box models (e.g., TreeSHAP and Integrated Gradients). Gunnarson et al. [12] carried out a systematic review of AI and ML methods for predicting realized and implied volatility indices. They concluded that these methods are highly promising and often achieve results equal to or better than those of traditional econometric models, with neural networks that use memory mechanisms such as LSTM and GRU among the best performers. Nevertheless, even with these findings, very few studies use XAI to support their empirical results, pointing to a major gap: as more complex models become increasingly common in

volatility forecasting, their decision-making processes remain largely unexplained.

Černevičienė and Kabašinskas [33] divided the FinXAI methodological landscape into three dimensions: the time at which the explanation is given (with inherently interpretable models such as LassoNet and decision trees being contrasted against post-hoc methods like SHAP, LIME, and Grad-CAM, which are used with pre-trained models), the scope of the explanation (this being either global, which shows the overall decision logic, or local, which explains a single prediction), and the applicability (with methods that are specific to the model and depend on its internal architecture, for example those based on attention mechanisms, versus model-agnostic methods that can be used with any black-box model).

In financial time series, feature selection helps identify relevant market indicators, lagged returns, volatility measures, macroeconomic variables, or technical indicators that contribute to forecasting price movement, risk, or market regimes. Industry portfolio indices thus gain intrinsic interpretability without sacrificing the ability to model non-linear relationships. Giudici et al. [38] emphasized that time-series XAI must account for temporal dependence, since feature importance changes over time and static explanation methods can lead to incorrect conclusions. Mohamed et al. [34] went beyond applying XAI to regression-based forecasting and used it for pattern-recognition tasks, predicting five different kinds of market trends. Sharbati et al. [35] combined LSTM (to capture temporal dependencies), XGBoost (for non-linear regression), and SHAP (for explanation) using Transformer-style encoders in a hybrid framework that balances predictive accuracy and transparency. Suprihadi et al. [36] applied a similar LSTM-SHAP approach to the forecasting of REIT returns. Ranjan et al. [22] found that a simpler 1D-CNN architecture achieved both interpretability and competitive predictive performance compared with ARIMA and LSTM in tests on S&P 500 data, showing that the trade-off between complexity and interpretability is not necessarily irreducible. Freeborough and van Zyl [37] showed that existing XAI methods do not transfer easily to RNN-based financial forecasting and that temporal dependencies require special handling.

Taken together, the literature indicates that the challenges encountered in XAI for financial time series differ from those in tabular or image-based applications. Because of autocorrelation and temporal dependence, explanations must capture dynamic patterns rather than static feature-output relationships, and because the data are non-stationary, predictor importance can vary across market regimes. The low

signal-to-noise ratio also raises the possibility that explanation techniques will assign importance to noise, and mixed-frequency data—such as daily market prices alongside monthly macroeconomic indicators—make it difficult to build a consistent explanatory framework. Furthermore, most XAI methods offer only correlational rather than causal explanations, which limits their value for financial decision-making. Current approaches include intrinsically interpretable models like LassoNet, post-hoc methods such as SHAP, and hybrid architectures such as LSTM-XGBoost, all aiming to balance predictive accuracy and transparency. Nevertheless, temporal dependence, regime shifts, and high noise remain major challenges. XAI therefore needs to develop explanation methods specifically designed for time-dependent data and extend evaluation across a wider range of financial contexts, market regimes, and asset classes before it can be considered a reliable part of financial decision-making.

2.2 Intrinsic Interpretability vs. Post-Hoc Explainability: An Ongoing Debate

Within the area of Explainable Artificial Intelligence (XAI) there exists a basic division concerning the way in which interpretability should be achieved. On one side, there is the view that intrinsically interpretable models should be used—meaning models in which interpretability is built right from the start—while on the other side it is argued that post-hoc explanation methods should be employed, involving the initial training of a high-performance "black-box" model and then the application of various tools to explain the decisions of that model. Madsen et al. [38] describe these as two distinct approaches: the intrinsic approach, which holds that only models intentionally designed to be interpretable can be interpreted, and the post-hoc approach, which asserts that black-box models can also be explained. This section examines the main arguments, empirical evidence, and the developing consensus, focusing on their relevance to significant financial decisions. It concludes by examining methods that go beyond this simple dichotomy.

The idea of intrinsic interpretability is to build interpretability directly into the model from the start, rather than adding it on afterwards. Zschech et al. [39] refer to this as "inherently interpretable machine learning", meaning that the decision logic can be understood on the basis of the architecture itself and not with the aid of any external tools. Models in this category include linear and logistic regression, decision trees, rule lists, Generalized Additive Models, and LassoNet. Shu et al. [40] demonstrate that an intrinsically interpretable model for predicting the shear strength of reinforced concrete beams can achieve accuracy comparable to black-box methods.

The main benefit of intrinsic models is their faithfulness. Because the decision logic is built directly into the model's structure, there is no need to translate or approximate the model's behavior, and the explanation is complete rather than only local. Yet this disadvantage shows itself when models are employed for complicated tasks such as forecasting financial time series, since simple models generally fail to capture non-linear patterns and long-range dependencies, resulting in the usual compromise between accuracy and interpretability. However, Rudin [26] challenges this view, arguing that the accuracy-interpretability trade-off is often a false dilemma; in many high-stakes situations, carefully designed interpretable models can match the accuracy of black-box models. Her main point is that researchers should develop inherently interpretable models rather than generating explanations afterward, especially since unreliable post hoc explanations in fields such as medical diagnosis, criminal justice, or financial risk management can have serious consequences.

Post-hoc explainability is based on the idea of "predict first, then explain": once a high-performance black-box model has been used, a variety of explanation techniques can show how the model reaches its conclusions. This includes methods like feature attribution (for example, SHAP, LIME, and Integrated Gradients), gradient- and propagation-based techniques such as Saliency Maps [41] and Grad-CAM [42], surrogate models [43] which either locally or globally approximate the black box, counterfactual explanations, and rule extraction. The main benefit of these methods is their flexibility: they can be combined with any black-box model, allowing researchers to use deep learning representational power for tasks such as financial forecasting without sacrificing accuracy at the design stage, at least in theory.

Another limitation is that feature attribution should not be taken as implying causation. Fu et al. [44] show this difference by combining explainable AI with causal inference techniques in an investigation of environmental drivers, demonstrating the advantage of supplementing attribution-based explanations with approaches that examine causal relationships. This point is especially important in high-stakes applications, since a predictor that has a significant influence and is identified by an XAI method does not have to be seen as one whose change would alter the predicted outcome.

The second point is instability, since minor alterations to the input can result in large changes to the explanation output. Thirdly, it is possible to generate spurious or unjustified explanations [27], which

demonstrate that counterfactual explanations can be disconnected from the true data distribution and cannot be derived from the actual data points by a continuous path; the authors regard this risk as 'substantial' in a number of datasets and stress that it actually leads to the creation of mathematically plausible but meaningless 'phantom explanations'. Fourthly, evaluation is difficult: although Munoz et al. [45] have suggested six explainer-independent metrics for measuring explainability, the need to propose them shows that the community has not yet reached consensus on evaluation standards.

The situation is especially serious when the stakes are high. Laugel et al. [27] argue that black-box models, even with post-hoc explanations, should be avoided when decisions affect people's lives, freedom, or financial security, since such explanations might not accurately reflect how the model behaves, could fail to show how it goes wrong, or might reveal spurious correlations. As a result, auditability and accountability weaken in such sensitive applications. Mersha et al. [46] make the same point, emphasizing that XAI should promote transparency, accountability, and fairness rather than simply serving as a technical add-on. Yet empirical evidence shows that neither intrinsic nor post-hoc methods are always the better option. Ai and Ramasamy [47] find that intrinsic explanations are generally more informative and reliable for product search. In contrast, Madathil et al. [48] think intrinsic methods are better for assessing process robustness in smart manufacturing, while post hoc methods better explain complex nonlinear relationships. Chakraborty et al. [49] come to the same conclusion, saying the right approach depends on the specific task and context.

They also discuss the alleged trade-off between predictive accuracy and interpretability. Laugel et al. [27] argue that this compromise is generally overemphasized and mainly stems from researchers' preference for black-box models rather than from any inherent flaw in interpretable models. Meanwhile, critics note that areas such as financial time series involve nonlinear dynamics, temporal dependence, and long-range interactions that simpler models may struggle to handle. In such cases, performance gains from architectures like LSTMs and Transformers could be real. The key point is not whether interpretability or accuracy should always take priority, but how to balance them given the application's specific requirements. In fields that are regulated, such as credit scoring, transparency may be essential, while in quantitative trading, even small improvements in predictive performance can have significant economic value [23, 25, 50].

As a result, the debate between intrinsic and post-hoc interpretability should not be seen as a simple choice between transparency and accuracy. A more useful question is whether the extra predictive power of a more complex model warrants the associated loss of direct interpretability and the extra uncertainty introduced by post hoc explanations. This matters especially in credit risk assessment, where transparency, justification, and auditability are essential to the decision-making process, even though more sophisticated nonlinear models can still provide predictive benefits. The growing body of evidence thus supports a comparative approach rather than making a priori assumptions: intrinsically interpretable and post-hoc explainable models should be compared not only on the basis of their predictive performance but also with regard to the nature, complexity, and reliability of the explanations they offer [23, 25, 33, 50, 51]. This kind of approach goes beyond the idea that greater model complexity automatically means either better predictive performance or worse practical interpretability and instead treats the relationship between predictive ability and interpretability as an empirical issue.

2.3 Benchmarking and Metrics for Explanation Quality

A major gap in the explainable AI (XAI) literature is the lack of standardized benchmarks for comparing explanation techniques. To address this gap, several recent efforts have emerged. Chakraborty et al. [49] developed a comprehensive benchmarking framework used to study both intrinsic and post hoc XAI methods on three typical 3D datasets. They evaluate explanations quantitatively using the Correctness (AOPC), Completeness (AUPC), and Compactness criteria, applying them consistently across datasets. Most importantly, the results indicate that explanation quality varies greatly across methods and domains, with no single approach performing best across all domains, underscoring the need for multi-metric evaluation. Statistical tests (Kruskal–Wallis and Mann–Whitney U) show significant differences in method performance. This aligns with the approach adopted in the present study, which assesses the explainers using three complementary metrics—faithfulness, stability, and consistency—rather than relying on a single criterion.

Metsch and Hauschild [52] have developed BenchXAI, a new package for benchmarking XAI methods that enables detailed evaluation of 15 different XAI techniques on biomedical tasks. Their evaluation found that Integrated Gradients, DeepLift, DeepLiftShap, and GradientShap performed well on all three tasks, while methods like Deconvolution and Guided Backpropagation did not. This again shows

that XAI performance varies by task and that systematic benchmarking is essential.

Bodria et al. [53] apply this benchmarking approach to tabular, image, and text data and express the quantitative comparisons in terms of faithfulness, stability, robustness, and computational time. Their experiments with tabular data are especially pertinent to assessing feature-attribution methods. Although the local surrogate methods generally achieved high fidelity in approximating the underlying black-box models, the findings on faithfulness and stability differed considerably. SHAP showed greater faithfulness than LIME, and none of the methods examined was consistently stable when inputs changed slightly. The authors note that no existing objective metric can identify a single best explainer, since different metrics reflect different desirable characteristics. This difference matters because an explanation can accurately approximate the model's outputs yet remain unstable under small perturbations, showing that fidelity and stability measure related but not equivalent aspects of explanation quality.

Belaid et al. [54] have proposed Compare-xAI, a benchmark that combines a variety of functional testing approaches for post hoc XAI algorithms. This benchmark illustrates the difficulty of evaluating XAI methods through a hierarchical scoring system designed for three groups of end users: researchers, practitioners, and the general public. The tests are divided into five categories—fidelity, fragility, stability, simplicity, and stress tests—with an overall comprehensibility score at the top level. This multi-dimensional evaluation method aligns with the present study's use of complementary metrics. Bragança and Souto [55] have introduced the Unified Time Series Classification Framework for Explainable Artificial Intelligence (UTS-XAI), which combines a traditional classification pipeline with comprehensive, time-aware XAI evaluation metrics such as faithfulness, robustness, sensitivity, stability, and localization. The framework supports several explanation methods—specifically SHAP, LIME, and Saliency Maps—and offers a way to systematically compare them using adapted versions of well-established XAI metrics reinterpreted for temporal data.

Consequently, explanation evaluation has gradually shifted from focusing on specifying individual desirable properties to developing systematic benchmarking procedures. The faithfulness criterion used in this study follows the approach of Samek et al. [56] which assesses whether the model's output deteriorates as expected when important features are removed. The stability criterion is based on the principle put forward by Alvarez-Melis and Jaakkola

[57] that similar inputs should yield similar explanations; their analysis shows that many commonly used interpretability methods violate this requirement. Lastly, consistency assesses whether independently generated explanations provide compatible descriptions of the same model behavior, drawing on the disagreement framework introduced by Krishna et al. [58] and further supported by empirical work by Brughmans et al. [59].

Building on this foundation, explanation quality is defined by measurable criteria for faithfulness, stability, and consistency, and by specific acceptance thresholds based on empirical research on the latest methods. The suggested framework combines quantitative assessments with user-centered criteria, enabling both researchers and practitioners to judge whether an explanation meets the established standards. This formalization is particularly relevant to the present study, which takes faithfulness, stability, and consistency as its three main evaluation dimensions.

3 Data and Methodology

Fig. 1 summarizes the study design. Two features distinguish this work from existing explainer benchmarks: the same-model contrast in stage 3, which holds the trained network fixed while varying the explanation, and the decomposition of stability into three perturbations in stage 5, which holds the explanation fixed while varying what it must withstand.

Five-day realized volatility is forecast for 20 equities across three markets (stage 1) from two feature representations chosen to contrast correlation structure rather than width—the nominally narrower panel spans more effective dimensions (stage 2). We estimate five model families under a purged rolling walk-forward split, each paired with the explainer appropriate to its architecture (stage 3). One arm is treated differently: we evaluate a single trained TabNet network using both its intrinsic attention mask and post hoc Integrated Gradients. Hence, any divergence between the two is attributable to the explanation method rather than to the model, a premise verified by checking that no attribution call alters the network's state. A cell enters the interpretability analysis only if it improves on step-5 persistence out-of-sample (stage 4). We then evaluate admitted explanations for faithfulness against a random-order control, stability under three perturbations that differ in what each fixes, and agreement between two explainers of the same network, with all comparisons made using matched-stock paired inference (stage 5).

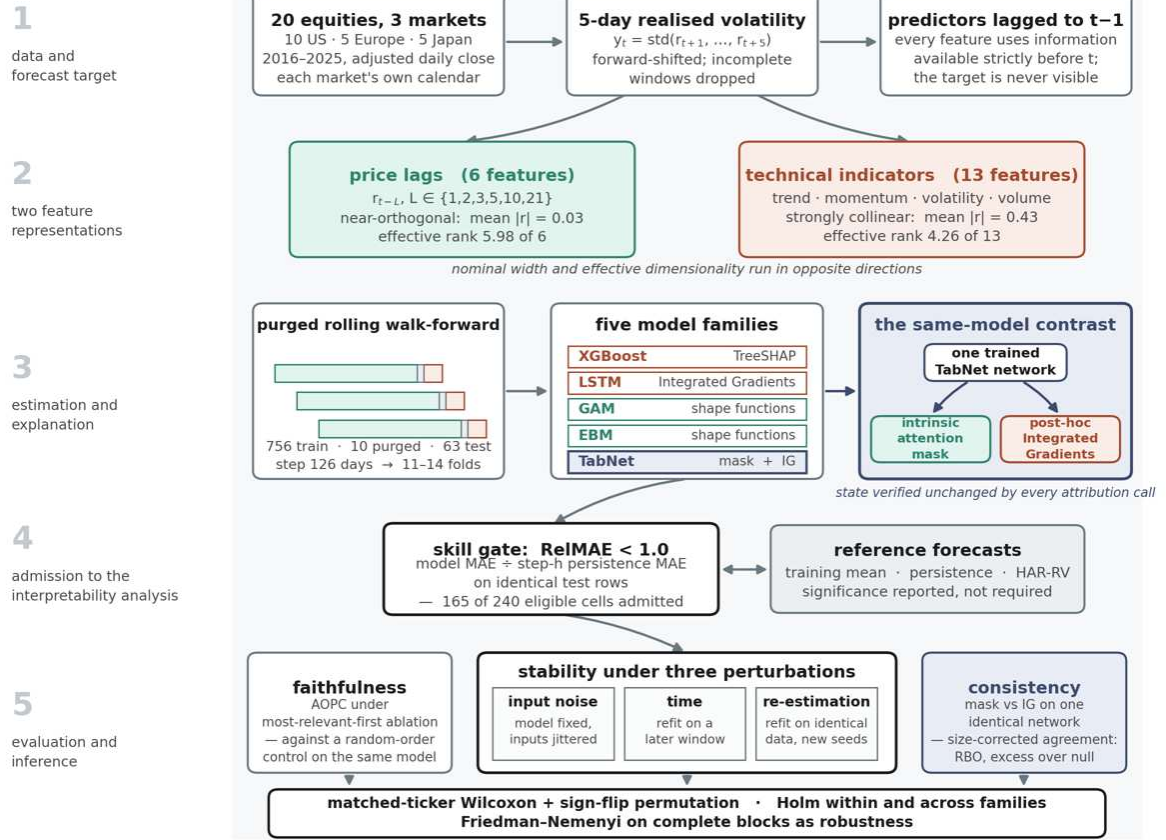


Fig. 1 Study design

3.1 Task and Data

We forecast 5-day-ahead realized volatility for 20 equities spanning three markets: 10 US large-caps (AAPL, MSFT, JPM, XOM, PG, JNJ, WMT, NVDA, KO, CAT), 5 European blue-chips (SAP.DE, MC.PA, SIE.DE, NESN.SW, ASML.AS), and 5 Japanese large-caps (7203.T, 6758.T, 9984.T, 8306.T, 6861.T), using adjusted daily close prices from 2016-01-01 to 2025-12-31. Each market's own trading calendar is used as-is — we do not force a common calendar or forward-fill one market's holidays using another's. Because the universe spans three markets with different trading calendars, liquidity profiles, and volatility regimes, we also report every primary interpretability metric by region as a robustness check (Section 5.7) to verify that pooled-sample findings are not driven by a single market's idiosyncrasies.

Let r_t denote the daily log return. The forecast target at time t is the realized volatility over the following $h = 5$ trading days,

$$r_t = \ln\left(\frac{P_t}{P_{t-1}}\right) \quad (1)$$

$$y_t = \text{std}(r_{t+1}, \dots, r_{t+h}) \quad (2)$$

computed by forward shifting so that y_t never uses information available only after t ; rows with an incomplete forward window are dropped. We chose volatility over raw returns as the target because daily returns are close to unpredictable under the efficient-market hypothesis, which would leave every model failing the skill gate and the interpretability comparison vacuous — realized volatility has genuine short-horizon persistence and is a standard target in the risk-management literature.

3.2 Feature Sets

Every feature at time t uses information only through $t-1$; the target is never observable during feature construction. We compare two feature representations, evaluated as separate experimental arms, so we can attribute shifts in explainer behavior to the input representation rather than conflating them with changes in the target or model: six price lags and thirteen technical indicators. Table 1 gives the exact formula for every feature as implemented, where p = adjusted close shifted 1 day; r = log return shifted 1 day; v = trading volume shifted 1 day; h, l = high/low shifted 1 day — all enforcing the $t-1$ information rule.

Table 1: List of price lags and technical indicators

Feature	Group	Formula
$r_{\text{lag}\{1,2,3,5,10,21\}}$	price_lags	$r_{(t-L)}$ for $L \in \{1,2,3,5,10,21\}$
$\text{sma}_5 / \text{sma}_{10} / \text{sma}_{21}$	Trend	$p_{t-1} / \text{mean_N}(p) - 1$ (price relative to its N-day mean)
$\text{ema}_{12} / \text{ema}_{26}$	Trend	$p_{t-1} / \text{EMA_N}(p) - 1$ (price relative to its N-day exponential mean)
rsi_{14}	Momentum	$100 \times \text{avg_gain}_{14} / (\text{avg_gain}_{14} + \text{avg_loss}_{14})$, Wilder-smoothed ($\alpha = 1/14$) gain/loss of r
macd	Momentum	$[\text{EMA}_{12}(p) - \text{EMA}_{26}(p)] / p_{t-1}$
roc_{10}	Momentum	$p_{(t-1)} / p_{(t-11)} - 1$
atr_{14}	Volatility	Wilder-smoothed 14-day average of $\max(h-l, h-p_{\text{prev}} , l-p_{\text{prev}})$, divided by p_{t-1}
bb_width	Volatility	$4\sigma_{21}(p) / \text{mean}_{21}(p)$ (standard $k=2$ Bollinger bands)
ret_std_{21}	Volatility	$\sigma_{21}(r)$
obv	Volume	252-day z-score of the 21-day mean of $\text{sign}(r) \cdot v$ (signed volume flow)
vol_z	Volume	$(v - \text{mean}_{21}(v)) / \sigma_{21}(v)$

The technical indicator (technical) panel is presented on a scale-free basis; the moving averages are shown as the ratio of price to the average rather than the average itself. The MACD and ATR are adjusted to the price at the same time, and signed volume flow is used as a rolling standardized score rather than a cumulative total over the whole sample. This method matters because it standardizes features within each training window. Such representations are well established in the literature on return and volatility prediction and are not arbitrary choices. In large-scale cross-sectional studies—for example, those carried out by Krauss et al. [60]—lagged-return panels are used as the only predictor set, and in the LSTM sequence models of Fischer and Krauss [61] as well, both trained on the entire S&P 500 cross-section. Technical indicator panels are among the main features in broader machine-learning asset-pricing studies, such as those of Gu et al. [62]. In the present study, Christensen et al. [1] use machine learning to forecast realized volatility using lagged autoregressive predictors, similar to those used in the price-lags representation described here.

A precedent alone is not enough to explain why the two panels differ by about a factor of two in nominal width. This decision is deliberate and departs from the goal of maximizing forecasting accuracy. The main aim is to examine explainer behavior rather than attain state-of-the-art predictive performance. To this end, the two panels contrast a narrow, nearly orthogonal representation with a nominally wider but highly collinear one.

The following three limitations of this framework must be made clear. Firstly, the two panels vary not only in width and correlation structure but also in the types of features used: raw autoregressive lags versus engineered, mutually correlated indicators. Since these different aspects are mixed, none of the results below can be attributed solely to the number of features. Given that effective dimensionality is inversely related to nominal width, the comparison is always made in terms of the correlation structure rather than dimensionality. Second, the two panels are fed into separately trained models, so cross-panel comparisons reflect differences between two model–representation combinations rather than a single model assessed under two different data views. Only in the within-panel comparisons, where two explainers are applied to a single trained network, is the model fixed. Thirdly, the price-lag panel contains only signed return lags and excludes absolute returns, squared returns, or lagged realized variances. This omission is necessary because lagged realized variances are highly autocorrelated and would introduce the collinearity this part of the study is intended to avoid. As a result, the panel provides only limited direct information on volatility. As shown in Section 5, an HAR-RV benchmark based on the excluded lagged realized variances attains the highest median accuracy in the study. The design therefore sacrifices some predictive information to enable a controlled comparison of correlation structures, and it makes no assertion about state-of-the-art accuracy.

3.3 Models and Explanation Methods

We evaluate five model families and two post-hoc explainers, pairing each model with the explainer or explainers appropriate to its architecture. Three of the five—the GAM, the EBM, and TabNet—have intrinsic explanations, and only TabNet is additionally paired with a post-hoc explainer, making it the only architecture on which we can compare an intrinsic and a post-hoc account of an identical trained network. This is a deliberate scope restriction, not a claim that the additive models admit no post-hoc explanation: both the GAM and the EBM could be explained by kernel-based Shapley attribution of the fitted object, which would yield three same-model contrasts instead of one. We record the omission as a limitation in Section 6.7 and as an immediate extension in the future directions in Section 6.7.

3.3.1 Extreme Gradient Boosting (XGBoost)

XGBoost is a gradient-boosted decision-tree algorithm that sequentially constructs trees to minimize prediction error while controlling model complexity through regularization. In this study, XGBoost was configured with 300 trees, a learning rate of 0.03, subsampling of 0.8 for observations and features, L2 regularization, and a minimum child weight of 5. Histogram-based tree construction was used to improve computational efficiency.

3.3.2 Long Short-Term Memory (LSTM)

Long Short-Term Memory (LSTM) is a recurrent neural-network architecture designed to capture temporal dependencies through gated memory mechanisms. The model used in this study consisted of a single LSTM layer with 32 hidden units followed by a linear regression output layer. Models were trained for 100 epochs using Adam optimization with a learning rate of 10^{-3} and a batch size of 256. Input sequences were constructed using rolling 20-day windows from the standardized feature data.

3.3.3 Generalized Additive Model (GAM)

GAM is an intrinsic model of the form $g(\hat{y}) = \beta_0 + f_1(x_1) + f_2(x_2) + \dots + f_p(x_p)$, where each is a smooth, learnable shape function, often splines or regression trees, and is a link function. Because the model simply adds separate contributions from each feature, we can plot and inspect each feature's learned curve directly, with no additional explanation needed. This transparency is the hallmark of intrinsic explainability, but it doesn't automatically capture feature interactions. We built a regression model that draws a smooth, slightly flexible curve for every single input feature. We restricted the curves so they can bend a little bit (6 basis splines) but applied a mathematical penalty ($\lambda = 0.6$) to keep them from

getting too jagged. Because the model simply adds up these individual curves, we can look at a graph of any single feature and see exactly how it drives the prediction.

3.3.4 TabNet

TabNet is a neural-network architecture for tabular data that uses sequential attention to perform instance-specific feature selection. The model was configured with decision and attention dimensions of $n_d = n_a = 8$, three decision steps, and a relaxation parameter of $\gamma = 1.3$ and was trained for 50 epochs. We standardized the regression target during training and transformed it back to its original scale for prediction. TabNet serves as the primary same-model comparison in this study: we explained each trained network using both its intrinsic attention mask and post-hoc Integrated Gradients, enabling us to assess differences in explanation behavior while holding the predictive model fixed.

3.3.5 Explainable Boosting Machine (EBM)

EBM extends the GAM by using gradient boosting to learn each univariate shape function, and in its general form can also learn a small number of automatically selected pairwise interaction terms, giving $g(\hat{y}) = \beta_0 + \sum f_i(x_i) + \sum f_{ij}(x_i, x_j)$. In this study, interactions are disabled ('interactions = 0'), so the fitted model is purely additive: $g(\hat{y}) = \beta_0 + \sum f_i(x_i)$. Each component is directly visualizable as a curve, so the model remains intrinsically interpretable while often achieving accuracy competitive with black-box learners. We use a single outer bag to keep training time comparable across arms. Because it's purely additive, we can examine the trees for any single feature to see its global effect, and we can also break down any individual prediction to see exactly how much each feature contributed to that result.

3.3.6 TreeSHAP

TreeSHAP is a model-specific implementation of SHAP for tree-based models that computes feature contributions using the internal structure of the fitted tree ensemble. Based on Shapley values, it attributes each prediction to individual input features relative to a reference value. In this study, TreeSHAP was applied to the fitted XGBoost models to obtain feature-attribution vectors for subsequent faithfulness and stability analyses

3.3.7 Integrated Gradients (IG)

Integrated Gradients (IG) is a gradient-based post-hoc attribution method for differentiable models. It estimates feature contributions by integrating the gradient of the model output along a path from a reference baseline to the observed input. In this study, IG was applied to both the LSTM and TabNet models.

For the LSTM, absolute attributions were aggregated across samples and timesteps to obtain feature-level importance scores, whereas for TabNet, IG was computed directly on the tabular input features

3.4 Experimental Arms and Reference Forecasts

Table 2 summarizes the 12 explained experimental arms (E1-E12). Three unexplained reference forecasts accompany these arms and are assessed only for predictive accuracy; they are never assigned an explainer and are excluded from the interpretability

analysis: B0, the training-window mean; B1, step-5 persistence, which is the denominator of the skill gate and therefore has a relative error of exactly one by construction; and B2, HAR-RV, based on daily, weekly, and monthly realized-variance components. The fixed B1 value provides an end-to-end check that the skill gate numerator and denominator are calculated on identical test rows. Although we implement a GARCH(1,1) benchmark, we do not report it because it models conditional return variance rather than the h-day realized-volatility target and would require an explicit aggregation step to make the targets comparable.

Table 2. List of experiments for each model-explainer

Exp.	Feature set	Model	Explainer	Type
E1	price_lags	XGBoost	TreeSHAP	Post-hoc
E2	tech_ind	XGBoost	TreeSHAP	Post-hoc
E3	price_lags	LSTM	IG	Post-hoc
E4	tech_ind	LSTM	IG	Post-hoc
E5	price_lags	GAM	-	Intrinsic
E6	tech_ind	GAM	-	Intrinsic
E7	price_lags	TabNet	Mask	Intrinsic
E8	tech_ind	TabNet	Mask	Intrinsic
E9	price_lags	TabNet	IG	Post-hoc
E10	tech_ind	TabNet	IG	Post-hoc
E11	price_lags	EBM	-	Intrinsic
E12	tech_ind	EBM	-	Intrinsic

4 Experimental Design

4.1 Walk-Forward Evaluation Protocol

Accuracy and interpretability are evaluated using a purged rolling walk-forward split to prevent look-ahead leakage (Fig. 2). Each fold uses a fixed 756-trading-day training block (approximately three years), of which 746 rows are fitted after purging, a 63-day test window (approximately one quarter), and advances by 126 trading days. Because the target at time (t) depends on returns from (t+1) to (t+h), we remove the final (h+5=10) observations from each training block before fitting to ensure training labels do not extend into the test period. Without purging, five rows per fold, or approximately 0.7% of the 756-row block, would be contaminated. We estimate feature-standardization parameters (μ and σ) only from the purged training window and apply them to both training and test data. This procedure yields 14 folds per ticker for the price-lag and HAR feature sets and 12 for the technical-indicator set for US and European firms. For Japanese firms, the corresponding numbers are 13 and 11 because of the

shorter Tokyo trading calendar. We therefore report fold counts separately for each cell.

4.2 Accuracy Metrics and the Skill Gate

We evaluate predictive accuracy using RMSE, sMAPE, and relative mean absolute error (RelMAE) against a naive baseline computed on the same test observations. For the five-day volatility horizon, the baseline is step-(h)persistence, $\hat{y}_t = y_{t-h}$, rather than day-to-day persistence. Because adjacent five-day rolling volatility targets overlap substantially, a one-day persistence benchmark would have artificially low error. Matching the baseline lag to the forecast horizon therefore provides a more appropriate comparison. RelMAE is defined as the model’s out-of-sample MAE divided by the step-(h) persistence MAE on the same rows; thus, $\text{RelMAE} < 1$ indicates out-of-sample improvement over persistence. RelMAE is preferred to MASE because MASE scales out-of-sample error by an in-sample naive error, which does not directly establish whether the model beats the baseline out of sample and may yield different conclusions across volatility regimes.

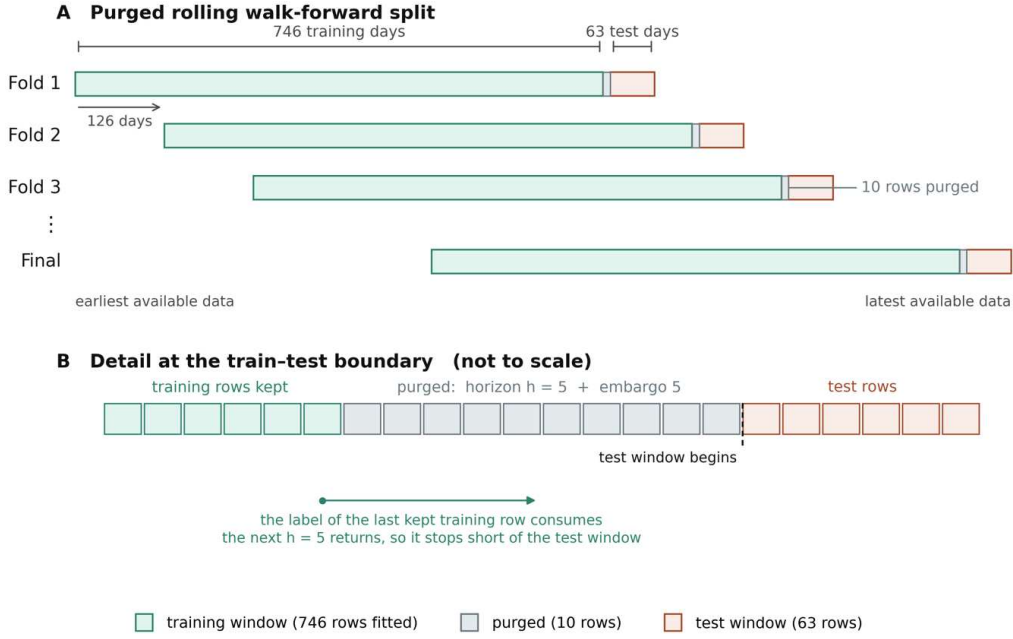


Fig. 2 Purged rolling walk-forward split

4.3 Explanation-Quality Metrics

We compute interpretability metrics for every cell, regardless of skill-gate admission, using each fold’s trained model and corresponding attribution vector. This lets us measure the skill gate’s influence on interpretability rankings directly. We calculate metrics at the fold level before averaging because using a single attribution vector averaged across folds would systematically disadvantage models whose explanations change over time.

4.3.1 Faithfulness

We measure faithfulness using area over the perturbation curve (AOPC) with most-relevant-first (MoRF) sequential ablation. We rank features by importance and cumulatively set them to zero, using the standardized training mean, starting with the most important feature. For the technical-indicator panel, we ablate entire collinear groups—trend, momentum, volatility, and volume—by group importance to avoid arbitrary rankings among near-duplicate indicators. RMSE is recalculated after each step. AOPC is the average RMSE increase over the top eight ablation steps, or over all six features for the price-lag panel; we also report a matched-depth result at ($k=5$). We additionally compare MoRF with reverse-order ablation and with a random-order control using the same model. Because a near-zero AOPC may indicate either an uninformative ranking or general model insensitivity to ablation, we interpret faithfulness relative to the random-order AOPC. We consider an

explanation more faithful only when its AOPC exceeds this control.

4.3.2 Stability

Three perturbation settings are evaluated. Input-noise stability keeps the fitted model fixed while adding Gaussian noise to the test data at $\sigma = 8\%$ of each feature’s standard deviation, repeated 20 times. We generate independent, reproducible random streams deterministically from the ticker, experiment identifier, and fold using a cryptographic hash. Temporal stability compares importance rankings across adjacent walk-forward folds after refitting the model, capturing changes caused by regime shifts and/or estimation noise. Seed stability refits the same model on identical data and folds using different random seeds, isolating variability attributable to the optimization path. We assess agreement under all three settings using Spearman rank correlation, rank-biased overlap (RBO), and top-(k) Jaccard overlap at both fixed and feature-scaled depths, interpreting results relative to expected agreement under independent random rankings.

4.3.3 Consistency

For TabNet, we compare intrinsic masks (E7, E8) with post-hoc Integrated Gradients (E9, E10) using the same trained model and the same agreement metrics applied to mean per-fold importance vectors for each ticker. Because E7/E9 and E8/E10 share identical fitted models, and attribution does not alter model parameters, observed disagreement reflects the explanation method rather than differences in predictive architecture.

4.3.4 Size-Corrected Agreement

Fixed-depth top-(k) overlap is not directly comparable across panels with different numbers of features. For example, with ($n=6$) and ($k=5$), the Jaccard index has a minimum of 0.667 and an expected random value of 0.7222, whereas for the 13-feature panel the expected value is 0.2537. To remove this panel-width effect, we additionally report Jaccard overlap at a feature-scaled depth, an excess-over-null normalization, $(\text{observed} - \text{expected}) / (1 - \text{expected})$, and RBO, which evaluates the full ranking with greater weight on top positions. Expected Jaccard values are derived analytically, whereas expected RBO is estimated by Monte Carlo simulation. Cross-panel conclusions are based on these size-invariant measures; raw fixed-depth values are retained only for comparison.

4.4 Statistical Testing

The main analysis uses a matched-ticker paired design in which each ticker contributes one value for each arm within a given metric and feature panel. For each pairwise comparison, we include only tickers with results for both arms, and we evaluate paired differences using the Wilcoxon signed-rank test and a two-sided sign-flip permutation test on the mean difference. We include the permutation test because it can be more reliable than the signed-rank test with small sample sizes. We report effect size using a matched-pairs rank-biserial correlation, which summarizes the balance of wins and losses across tickers. Pairing is essential because substantial ticker-specific variation could otherwise obscure differences attributable to the explainers, particularly when comparing intrinsic and post-hoc explanations derived from the same fitted network. Two secondary analyses assess robustness to incomplete overlap among admitted tickers. First, because different explainers may qualify on different ticker sets, we apply a Mann–Whitney U test to all admitted cells as an unpaired sensitivity analysis. Second, where sufficient complete blocks are available—that is, tickers for which all arms in a comparison group are simultaneously admitted—we use a Friedman test followed by Nemenyi post-hoc comparisons as an additional robustness check. We report the number of matched pairs for every comparison, and we classify analyses with fewer than five pairs as underpowered and do not interpret them as evidence of no difference. For pairwise explanation comparisons, we Holm-adjust p-values within each metric-feature-set family, then apply a second Holm correction across all families.

5 Results

5.1 Sample, splits, and the skill gate

The study contains 12 explained arms (E1-E12) plus three unexplained reference forecasts (B0-B2). Because E7/E9 and E8/E10 reuse the same fitted TabNet predictions for two different explainers, these correspond to 13 unique predictive configurations, or 260 stock-level fitted configurations across 20 equities, but 15 arm/reference result series, or 300 stock-level result rows. All predictive configurations trained without error. Walk-forward validation produced 14 folds per stock for the price-lag and HAR feature sets and 12 for the technical-indicator set on the US and European names, and 13 and 11 respectively on the five Japanese names, reflecting the shorter Tokyo trading calendar over 2016-2025. The technical-indicator arms lose two folds relative to their price-lag counterparts because the volume-flow normalization requires a 252-day warm-up; fold counts are therefore reported for each cell. Each training block is purged of the $h + \text{embargo} = 10$ terminal rows whose labels depend on returns inside the subsequent test window. Without the purge, five training rows per fold, approximately 0.7% of a 756-row block, would carry contaminated labels.

The skill gate admits a cell if its out-of-sample relative MAE against step-5 persistence, computed on identical test rows, falls below one. The gate admitted 165 of 240 eligible cells (68.8%) (Table 3). We report the significance of the per-cell skill advantage rather than require it: the gate is a screening step separating models with forecasting skill from those without, not a confirmatory test, and both the raw and Benjamini–Hochberg-adjusted one-sided fold-paired p-values accompany every cell. Requiring per-cell adjusted significance at $q \leq 0.05$ would admit only 6% of cells and would eliminate both same-model comparisons entirely; results under a stricter $q \leq 0.10$ tier are reported as a sensitivity analysis in 5.7. The persistence reference (B1) returns a relative MAE of exactly 1.000 on all 20 stocks by construction, serving as an end-to-end check that the gate's numerator and denominator align on the same rows.

Two patterns are visible. Every technical-indicator arm is admitted less often than its price-lag twin. The recurrent arm is the weakest in the study: IG/LSTM on technical indicators admits no stock at all. The gate admits US cells at roughly half the rate of the other two markets (Table 4). Because of where the gate is defined, the admitted sample is close to balanced across markets—59 US, 52 European, and 54 Japanese cells—whereas the universe was designed with ten US names against five in each of the other two regions.

Table 3. Cells admitted by the skill gate, 20 per arm.

Arm	Explainer / model	Features	Admitted
E5	GAM (intrinsic)	price lags	20
E1	TreeSHAP / XGBoost	price lags	18
E11	EBM (intrinsic)	price lags	16
E7	TabNet attention mask	price lags	16
E9	Integrated Gradients / TabNet	price lags	16
E2	TreeSHAP / XGBoost	technical	16
E12	EBM (intrinsic)	technical	14
E8	TabNet attention mask	technical	13
E10	Integrated Gradients / TabNet	technical	13
E3	Integrated Gradients / LSTM	price lags	12
E6	GAM (intrinsic)	technical	11
E4	Integrated Gradients / LSTM	technical	0

Table 4 Gate rate by market

Market	Names	Eligible cells	Admitted	Rate
Japan	5	60	54	90.0%
Europe	5	60	52	86.7%
United States	10	120	59	49.2%
All	20	240	165	68.8%

Fig. 3 shows relative mean absolute error against step-5 persistence for every stock and model-explainer arm, computed on identical test rows. Values below 1.00 improve on persistence; a heavy outline marks the 165 of 240 eligible cells admitted to the interpretability analysis. Stocks are grouped by market, and arms sharing one trained network appear once.

5.2 Forecast accuracy

Table 5 orders every arm and reference by median relative MAE. Values below one indicate an out-of-sample improvement in persistence. Three observations bear directly on interpreting the results. First, the HAR-RV benchmark achieves the best median accuracy in the study, but it separates only from some of the explained arms. In matched-stock paired comparisons across the 20 stocks (Table 6), HAR-RV significantly outperforms the EBM, both XGBoost arms, and TabNet, beating them on all 20 stocks. Against the GAM on price lags, it is

statistically indistinguishable: HAR-RV has a median paired difference of 0.004 in its favor, winning on 11 of 20 stocks, rank-biserial 0.12, $p = 0.648$. A three-parameter regression on realized-variance lags therefore matches the best-performing learned model in this study while decisively beating the neural and tree-based arms.

Fig. 4 shows the distribution of relative errors across the 20 stocks, one point per stock. The dashed line is the HAR-RV median and the dotted line the training-window mean; only the spline GAM on price lags is statistically indistinguishable from HAR-RV under matched-stock paired testing. Filled points are admitted by the skill gate. Table 6 shows HAR-RV against every explained arm, matched by stock ($n = 20$). Positive differences favor HAR-RV. We show arms sharing one trained network once, since their accuracy is identical by construction. p -values are Holm-adjusted across the ten comparisons. The median difference is the median of the paired differences and need not equal the difference of the two median columns.

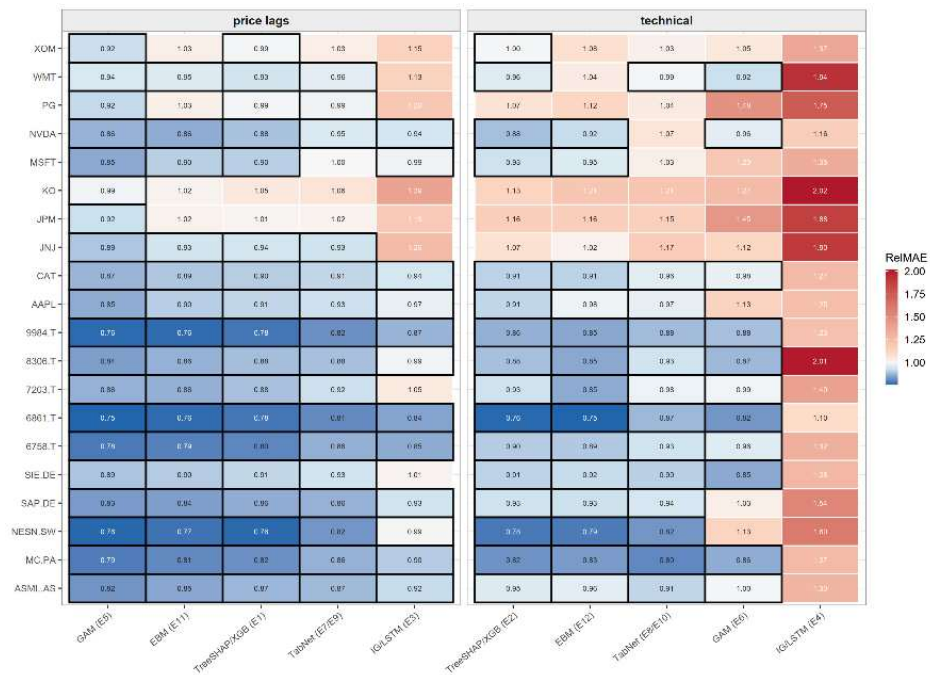


Fig. 3 Skill-gate heatmap

Table 5 Forecast accuracy, relative MAE against step-5 persistence on identical test rows.

Arm	Model	Features	Mean	Median	Min	Max
B2	HAR-RV	RV lags	0.852	0.844	0.758	0.984
E5	GAM	price lags	0.855	0.858	0.754	0.987
E11	EBM	price lags	0.887	0.877	0.762	1.030
E1	XGBoost	price lags	0.892	0.892	0.760	1.053
B0	Training mean	—	0.917	0.892	0.762	1.126
E2	XGBoost	technical	0.937	0.920	0.758	1.159
E12	EBM	technical	0.950	0.923	0.754	1.205
E7 / E9	TabNet	price lags	0.921	0.925	0.808	1.065
E8 / E10	TabNet	technical	0.979	0.965	0.805	1.211
E3	LSTM	price lags	1.025	0.987	0.842	1.389
E6	GAM	technical	1.048	0.992	0.824	1.480
B1	Persistence	—	1.000	1.000	1.000	1.000
E4	LSTM	technical	1.496	1.356	1.096	2.017

Table 6 HAR-RV against each explained arm

Arm	Model	Median RelMAE	Median diff.	HAR wins	Rank-bis.	p
E5	GAM / price lags	0.858	+0.004	11 / 20	0.124	0.648
E11	EBM / price lags	0.877	+0.034	16 / 20	0.810	0.0007
E1	XGBoost / price lags	0.892	+0.031	18 / 20	0.933	< 0.0001
E2	XGBoost / technical	0.920	+0.080	19 / 20	0.971	< 0.0001
E12	EBM / technical	0.923	+0.080	19 / 20	0.971	< 0.0001
E7/E9	TabNet / price lags	0.925	+0.064	20 / 20	1.000	< 0.0001
E8/E10	TabNet / technical	0.965	+0.119	19 / 20	0.981	< 0.0001
E3	LSTM – IG / price lags	0.987	+0.135	20 / 20	1	< 0.0001
E6	GAM / technical	0.992	+0.162	19 / 20	0.952	< 0.0001
E4	LSTM – IG / technical	1.356	+0.5181	20 / 20	1	< 0.0001

The following two interpretations are offered. The first decision mentioned in Section 3.2: the price-lag panel restricts the machine-learning models, based on a design includes only the signed return lags and omits the

realized-variance predictors used by HAR-RV to maintain the near-orthogonality the panel was intended to have. For this reason, we do not claim that our models achieve state-of-the-art accuracy; we include the gate to show that the models we compare have clear forecasting skill, not that they are the best forecasters available. The second interpretation is that the additive model, which is interpretable by construction, did not suffer any loss in

accuracy when compared either with the most powerful econometric benchmark or with any of the black-box competitors — a result that is in line with Rudin’s [26] argument and one which stands in sharp contrast to the other architecture in this study that is described as being interpretable by design, since HAR-RV outperforms it on all 20 stocks and ranks last in terms of explanation reproducibility (Section 5.5.3).

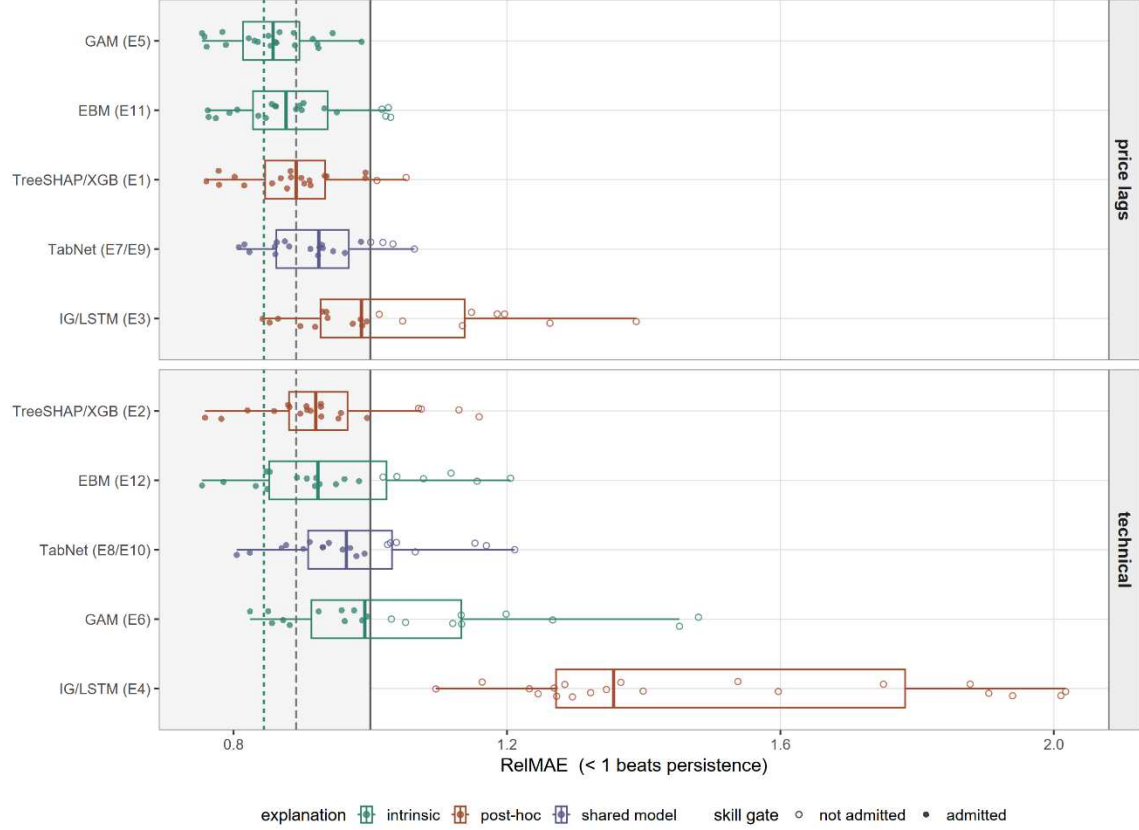


Fig. 4 Distribution of relative errors across the 20 stocks

Forecast difficulty varies considerably by market and is the same across arms. In Europe the median relative error among eligible cells is 0.882, and in Japan it is 0.870 as against 1.001 in the United States, meaning that the median eligible US cell shows no improvement in terms of persistence at all — which is the reason why the US gate rate is about one-half rather than indicating a general failure of the US arms. This difference holds for each of the eleven admitted

arms, with gaps ranging from 0.07 to 0.28, showing a market-level effect rather than an interaction between the market and the model family.

5.3 Structure of the two feature representations

Because several results below depend on the difference between the two feature sets, we first characterize that difference directly rather than by feature count (Table 7).

Table 7 Correlation structure, averaged over the 20 stocks.

Representation	Features	Mean $ r $	Max $ r $	Pairs $ r > 0.8$	Condition number	Effective rank
Price lags	6	0.029	0.071	0%	1.13	5.98
Technical	13	0.427	0.969	23%	120.5	4.26

The price-lag representation is effectively orthogonal since no pair of features has a correlation exceeding $|r| = 0.071$ and its effective rank—calculated as the exponentiated entropy of the normalized singular

spectrum—is 5.98 out of a possible 6. In contrast, the technical-indicator representation is highly collinear since about a quarter of the feature pairs have $|r| > 0.8$; the condition number is two orders of magnitude greater.

The opposite is true for the two representations with regard to their nominal dimensionality: the arm with more columns has fewer effective dimensions. A comparison between them should therefore be regarded as a comparison of feature type and correlation structure, not one of dimensionality, and Section 5.6 is framed on that basis.

5.4 Faithfulness

For each fold, we computed ablation-based faithfulness by removing features cumulatively in the order given by that fold's attribution vector. On the technical-indicator panel, where features are strongly collinear, the order is constrained so that members of the same cluster (trend, momentum, volatility, volume) are removed consecutively, clusters being visited in decreasing order of their summed importance; features are still removed individually, so the grouping constrains the order rather than the granularity, and an explainer is neither credited nor

penalized for an arbitrary tie-break among near-duplicate indicators. We apply no grouping to the price-lag panel, whose features are near-orthogonal. With each AOPC, we also give the AOPC achieved by carrying out the ablation in random order on the same model and fold, to tell the difference between 'the ranking is uninformative' and 'ablation has no effect on this model'. For ten of the eleven admitted arms, the importance ordering performs no better than a random ordering: AOPC minus its random-order control ranges from -0.00014 to $+0.00012$ (Table 8). The eleventh is the sole exception and points the wrong way: GAM on technical indicators yields an AOPC of 0.0055 against a random-order control of 0.0119, meaning that random ablation degrades that model substantially more than importance-ordered ablation does. For all four TabNet arms, the AOPC itself is negative, so removing features in the order the explanation ranks them slightly improves out-of-sample error.

Table 8 Faithfulness by arm, admitted cells

Arm	Model / Features	n	AOPC	Random Order	Difference
E5	GAM / price lags	20	0.00038	0.00032	+0.00005
E1	TreeSHAP / XGB / price lags	18	0.00030	0.00025	+0.00005
E11	EBM / price lags	16	0.00051	0.00041	+0.00011
E7	TabNet-mask / price lags	16	-0.00014	-0.00012	-0.00001
E9	IG / TabNet / price lags	16	-0.00015	-0.00011	-0.00005
E3	IG/LSTM / price lags	12	0.00025	0.00039	-0.00014
E2	TreeSHAP/XGB / technical	16	0.00009	-0.00003	+0.00012
E12	EBM · technical	14	0.00017	0.00005	+0.00012
E8	TabNet-mask · technical	13	-0.00031	-0.00030	-0.00002
E10	IG/TabNet · technical	13	-0.00031	-0.00023	-0.00008
E6	GAM · technical	11	0.00551	0.01190	-0.00639

The following two points arise. First, ablation-based faithfulness is not informative for this task, and we reach no different conclusion. The most likely explanation is the collinearity noted in 5.3 for the technical set, together with the near-orthogonality and very low signal-to-noise ratio of the price-lag set: when the target is so noisy, removing any one feature yields an out-of-sample RMSE increase that is no different from noise. Second, and importantly for interpreting the pairwise results, many pairwise AOPC differences among the arms are statistically significant. These effect sizes are on the order of 10^{-4} to 10^{-3} on a metric that does not exceed a random baseline. We include them for completeness, although we do not regard them as indicating faithfulness.

5.5 Stability of explanations under three perturbations

We alter each explanation in three different ways: we shift the estimation window into the future (this

is known as temporal stability), add Gaussian noise to the test inputs when using a fixed model (this is input stability), and re-estimate the model using new random seeds on the same folds (this is seed stability).

We assess agreement using Spearman rank correlation, rank-biased overlap, and a size-corrected top-k Jaccard; we perform inference with the matched-stock Wilcoxon signed-rank test with a sign-flip permutation test, and we Holm-correct both within each metric-feature-set family and across all families.

5.5.1 Temporal stability

Table 9 shows that the Temporal stability of both TabNet arms is considerably lower than that of all the other arms, whether in the different feature sets or when assessed using all the agreement measures, and all the comparisons between TabNet and the other arms remain significant after applying the Holm correction ($p \leq 0.01$).

The combined figures underestimate the TabNet result; when broken down by market, the intrinsic mask shows a temporal agreement of 0.003 in Europe and -0.020 in Japan for price-lagged values, and the post hoc Integrated Gradients on the same networks amount to -0.012 in Japan. A negative rank correlation indicates that the order of importance

between adjacent quarters is reversed, rather than simply uninformative. In two of the three markets, the features which TabNet ranks as most important in one quarter are generally the ones it ranks as least important in the following quarter. All the other arms have values between 0.39 and 0.69 in all markets.

Table 9 Temporal stability across adjacent folds, admitted cells

Arm	Explainer / model	Features	Spearman ρ	RBO	Top-k excess
E5	GAM	price lags	0.644	0.907	0.548
E1	TreeSHAP / XGBoost	price lags	0.554	0.861	0.356
E3	IG / LSTM	price lags	0.546	0.862	0.387
E11	EBM	price lags	0.478	0.848	0.274
E9	IG / TabNet	price lags	0.111	0.792	0.079
E7	TabNet mask	price lags	0.054	0.785	0.044
E2	TreeSHAP / XGBoost	technical	0.642	0.773	0.395
E6	GAM	technical	0.618	0.794	0.474
E12	EBM	technical	0.429	0.716	0.221
E10	IG / TabNet	technical	0.117	0.615	0.074
E8	TabNet mask	technical	0.073	0.597	0.045

The remaining arms are not statistically indistinguishable. GAM outperforms EBM in terms of price lags, as shown by the Spearman ρ (median difference 0.145, $p = 0.011$), on RBO (0.044, $p < 0.001$), and on top-k excess (0.246, $p = 0.011$), and it also performs better than TreeSHAP/XGBoost on RBO (0.041, $p = 0.001$), although it does not do so on ρ . It is therefore no longer justified to describe the non-TabNet arms as forming a single stable cluster; GAM in fact ranks first among the non-TabNet arms. When Integrated Gradients is applied to both the LSTM and to TabNet—the same attribution method used with two different architectures—the difference in Spearman ρ is 0.519 ($n = 11$, $p = 0.014$) and the difference in input stability is 0.076 ($p = 0.006$). For a single TabNet network, the two explanations obtained under the same perturbation are statistically indistinguishable (Section 5.5.4). Temporal stability depends on the model-explainer combination, not the attribution method alone. These results suggest that explanation stability can depend more strongly on the model–representation context than on the attribution method alone. Agreement between importance rankings of adjacent walk-forward folds, admitted cells only, under four measures. Diamonds mark arm means. Raw top-5 Jaccard is included for continuity with earlier work: at six features it can take only two values and is not comparable across panels (Fig. 5). The matched-stock comparisons of temporal stability between arms, by Wilcoxon signed-rank with a sign-flip permutation test, Holm-corrected within family, can be seen in Fig. 6. Color gives the rank-biserial effect size, red indicating that the row arm is the more stable; cells show the effect size and adjusted p . A

heavy outline marks the pair reading one identical network.

5.5.2 Input stability

In absolute terms, all arms are very robust to input jitter (ranging from 0.897 to 0.999), so the differences reflect only robustness differences. GAM shows the highest robustness across both feature sets, whereas IG/TabNet is the least robust. The order generally follows temporal stability, with one clear exception, which we discuss next.

5.5.3 Seed stability

We re-estimated the architecture five times on the last three folds of each stock using independently derived seeds, with all explainers attached to a configuration that reads the same seeded network; we report prediction and explanation agreements separately. Table 10 shows seed stability, where Pred. ρ is the mean pairwise correlation of prediction vectors across seeds, and Expl. ρ is the mean pairwise Spearman correlation of attribution vectors; CV is the coefficient of variation of fold RMSE across seeds.

GAM is exactly reproducible in both feature sets, as expected for penalized iteratively reweighted least squares, providing a validity check for the seed harness. The reproducibility of the predictions follows this order: GAM (0.000) < XGBoost (0.007–0.010) < EBM (0.014–0.038) < TabNet (0.062–0.091) < LSTM (0.073–0.159). Both neural architectures are considerably less reproducible than the tree-based and additive models, although which is worse depends on the feature set and the measure.

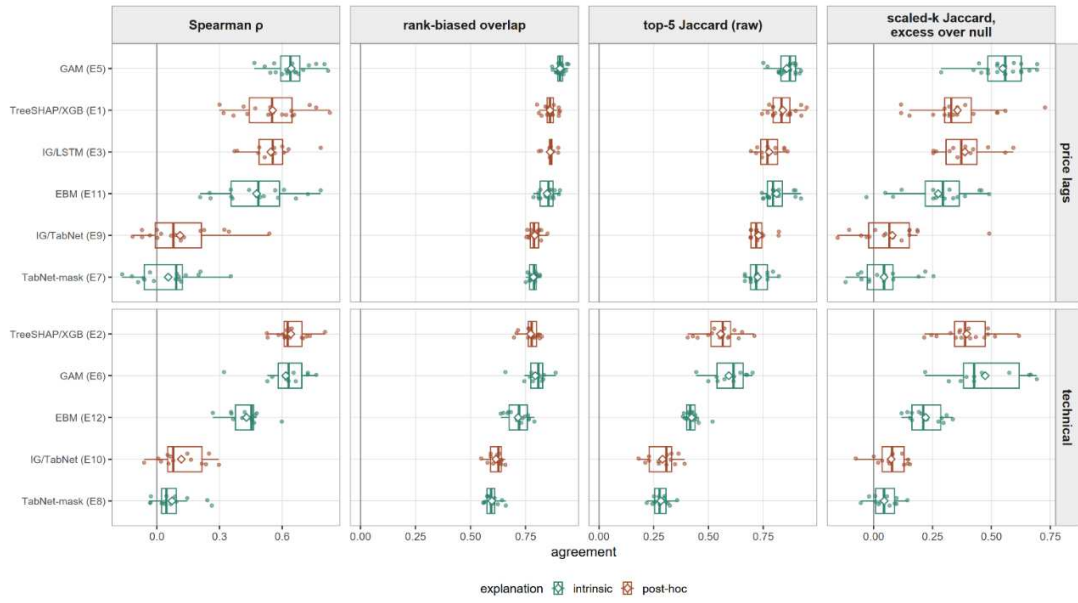


Fig. 5 Temporal stability

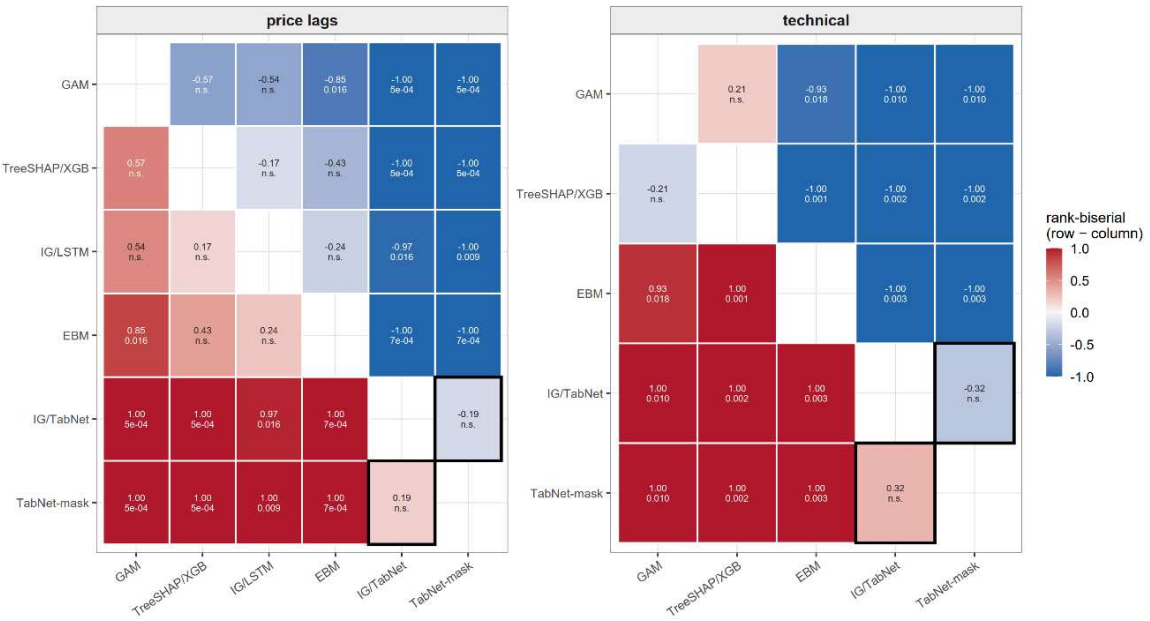


Fig. 6 Pairwise significance - matched-stock comparison

Table 10. Seed stability

Arm	Model	Features	Pred. ρ	Expl. ρ	Expl. – Pred.	CV
E5	GAM	price lags	1.000	1.000	0.000	0.000
E6	GAM	technical	1.000	1.000	0.000	0.000
E2	XGBoost	technical	0.966	0.947	−0.019	0.010
E1	XGBoost	price lags	0.958	0.921	−0.037	0.007
E12	EBM	technical	0.796	0.702	−0.095	0.038
E11	EBM	price lags	0.712	0.458	−0.253	0.014
E3	LSTM	price lags	0.399	0.494	+0.095	0.073
E10	IG / TabNet	technical	0.305	0.265	−0.040	0.091
E8	TabNet mask	technical	0.305	0.104	−0.201	0.091
E9	IG / TabNet	price lags	0.302	0.120	−0.182	0.062
E7	TabNet mask	price lags	0.302	0.044	−0.258	0.062
E4	LSTM	technical	0.190	0.493	+0.303	0.159

5.5.4 The same-model comparison, and a dissociation

The TabNet settings allow for the clearest contrast in the study. For each fold, we analyze a separately trained network using both an intrinsic attention mask and post hoc Integrated Gradients. Because both

approaches make identical predictions and agree at the prediction level, the pipeline checks this agreement across all 40 arm pairs. Table 11 shows Mask minus IG on the same trained network, matched by stock, with negative values favoring IG.

Table 11 Mask minus IG matched by stock

Perturbation	Measure	Features	n	Mask	IG	Median diff.	Rank-bis.	p
Input noise	Spearman ρ	price lags	16	0.967	0.897	+0.066	+1.000	< 0.0001
Input noise	Spearman ρ	technical	13	0.987	0.963	+0.026	+1.000	0.0002
Retraining	Spearman ρ	price lags	16	0.006	0.077	-0.072	-0.574	0.044
Retraining	Spearman ρ	technical	13	0.086	0.216	-0.135	-0.934	0.001
Retraining	RBO	price lags	16	0.782	0.790	-0.005	-0.574	0.044
Retraining	RBO	technical	13	0.599	0.652	-0.054	-0.890	0.002
Retraining	Top-k excess	price lags	16	-0.015	0.066	-0.037	-0.507	0.074
Retraining	Top-k excess	technical	13	0.048	0.130	-0.112	-0.912	0.002
Time	ρ , RBO, top- k	both	13–16	—	—	≈ 0	—	> 0.10
Faithfulness	AOPC, variants	both	13–16	—	—	≈ 0	—	> 0.27

When we apply the perturbation, the order of the two explanations reverses. The intrinsic mask proves to be uniformly more resistant to input jitter — it obtains the highest rank in every matched stock for both feature sets, resulting in a rank-biserial correlation of exactly +1.000 — and is considerably less robust to re-estimation, with the rank-biserial values ranging from -0.51 to -0.93 and showing the same direction on all three agreement measures. In the cases of temporal shift and ablation, the two explanations are indistinguishable. This counterpoint also deals with an objection to the input-stability result alone. Since Integrated Gradients is gradient-based, it is inherently sensitive to input noise, while the mask is selection-based; this definitional artifact would mean the mask is advantaged under all perturbations. It is not. In practice, the distinction between intrinsic and post-hoc approaches is not based on a single criterion. For explanations that must be robust to small input changes, the intrinsic mask is the better choice; whereas for global importance rankings that must withstand a model refresh, post-hoc attribution is preferable. Neither approach is superior.

One point deserves highlighting. The attention mask of TabNet—the element that makes the architecture interpretable by design—reaches a seed-to-seed rank agreement of 0.006 for price lags and 0.086 for technical indicators. Yet the same networks' predictions correlate at 0.27 and 0.25. The explanations change almost entirely when the network is retrained, and they change more than the function they are supposed to explain. There are three points to note. The top- k excess result on price lag is the only marginal entry in Table 10 ($p = 0.086$); the other seven have $p \leq 0.009$. For the seed contrast, all 20 stocks are included since seed stability is calculated independently of the gate, but for the

input- and temporal-stability contrasts only the admitted cells are used (with numbers ranging from 13 to 16. Lastly, this dissociation is based on a single architecture: TabNet is the only model in the study that provides both intrinsic and post-hoc explanations; this study has not yet determined whether this reflects a general distinction between intrinsic and post-hoc explanations or is specifically related to attention mechanisms.

5.6 Explainer Agreement and the Top-k Artifact

When comparing the two explanations produced by a single TabNet network using fixed-depth top-5 Jaccard scores, we observe greater agreement for the six-feature price-lag representation than for the 13-feature technical representation, with agreement of 0.854 versus 0.549 on the admitted cells (Table 12).

For any two subsets consisting of five elements taken from a set of six elements, the smallest number of common elements is four. As a result, the top-5 Jaccard score in the price-lag representation can take on only the values 0.667 or 1.000. The lowest possible value is 0.667, compared with 0.000 for the 13-feature set, and the value expected under independent random rankings is 0.7222, compared with 0.2537. Hence, this measure cannot capture disagreement when $n = 6$ and is not comparable between the two representations.

The per-stock values show this limitation. For XOM and 6758.T, the two explainers rank the six features in almost opposite orders (Spearman ρ of -0.143 and -0.257, respectively), yet the top-5 Jaccard score is still 0.667 for both. With 7203.T, the top-5 Jaccard is 1.000, although the Spearman ρ is only 0.143. Seven of the sixteen admitted price-lag cells end up below chance agreement after allowing for the random-

ranking floor, even though the raw measure is 0.667. Three of the four measures, each either size-invariant or explicitly floor-corrected, show greater agreement between explainers for the 13-feature representation. Although Spearman's ρ has an expected value of zero under the null hypothesis of random ranking at both dimensions and is not influenced by floor effects, it

previously showed the same direction of result in the earlier version of the study but was neglected because it was considered unstable. The only exception is the floor-corrected fixed-k Jaccard, which still favors price lags since the median value for that arm has reached the measure's

Table 12 Mask-versus-IG agreement by feature representation, admitted cells.

Measure	Price lags (n = 6)	Technical (n = 13)	Direction
Fixed top-5 Jaccard	0.854	0.549	price lags higher
Spearman ρ	0.464	0.633	technical higher
Size-scaled Jaccard, excess over null	0.228	0.325	technical higher
RBO, excess over null	0.330	0.381	technical higher
Fixed top-5 Jaccard, excess over null	0.460	0.391	price lags higher

ceiling of 1.000; this leads to a saturation effect, unlike the floor problem, which the fixed-depth top-k overlap cannot prevent when the sample size is small.

Agreement between the intrinsic attention mask and post-hoc Integrated Gradients read from one identical

network, by stock and feature panel. Raw top-5 Jaccard reports higher agreement on the six-feature panel, with a floor of 0.667 and a ceiling of 1.000; all three size-invariant measures reverse that ordering, and the bars show medians (Fig. 7).

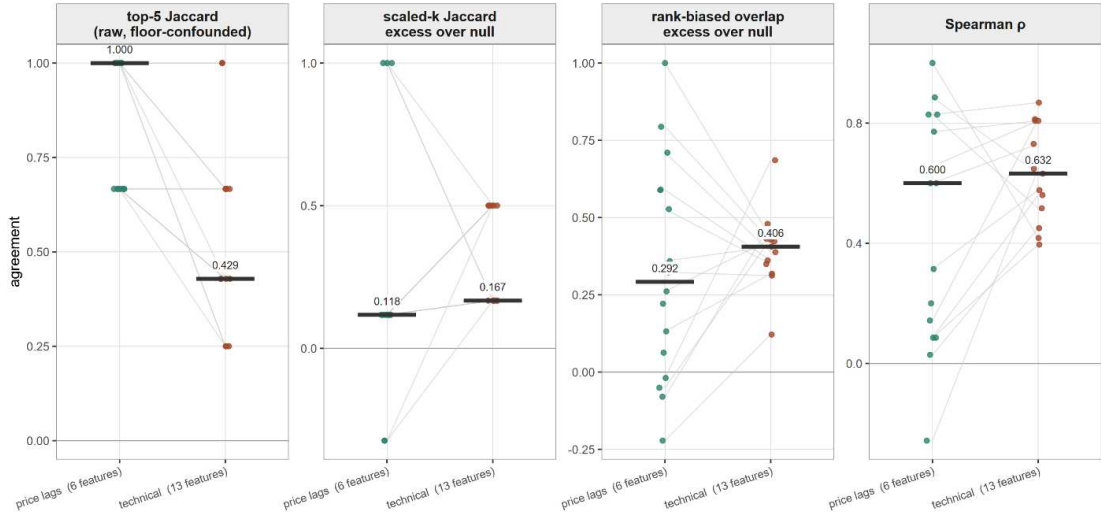


Fig. 7 Mask versus Integrated Gradients agreement

5.7 Robustness and Integrity Checks

The model is unaffected by attribution. The assumption stated in 5.5.4 is that both explanation types are based on one trained network. Each attribution request records the full model state dictionary, the parameter tensors, and the non-parameter buffers alike, both before and after execution. Across the entire study—covering LSTM Integrated Gradients, TabNet Integrated Gradients, and the TabNet attention masks—we observed no state changes. Specifically, TabNet Integrated Gradients runs in evaluation mode, so batch normalization reads the running statistics but does not update them. Therefore, the mask and the attribution refer to the same normalization state.

The gate does not produce the interpretability ranking; we calculated interpretability metrics for all cells, whether or not they were admitted, so we could directly compare the two groups in each arm. Of the 36 arm $\times$ metric tests, only one yields a p-value less than 0.05: AOPC for IG/LSTM on price lags (admitted median 0.0004, non-admitted -0.0004 , $p = 0.004$). Since about 1.8 significant results would be expected by chance, and the metric referred to in 5.4 carries no signal, we do not consider this finding. For any arm, no stability metric shows a difference between admitted and non-admitted cells.

Under a sensitivity tier that additionally requires a Benjamini–Hochberg-adjusted $q \leq 0.10$ on the per-cell skill test, 112 of 240 eligible cells are admitted (46.7%), against 165 (68.8%) under the primary gate.

Every arm retains at least six cells, and all four shared-model arms survive, so the same-model comparison of 5.5.4 remains available. The arm ordering is unchanged: ranking the arms by their mean value within each tier gives a Spearman correlation of 1.000 between the primary and strict tiers on temporal agreement, rank-biased overlap and input stability alike. The conclusions reported above therefore do not depend on where the skill gate is set. The tier is not neutral in composition, however: price-lag arms lose between one and six of their admitted cells while technical arms lose six or seven, so the stricter sample tilts further toward the price-lag panel — a further reason to prefer the screening gate as primary.

The robustness across regions. Since the universe involves three trading calendars, different liquidity profiles, and various volatility regimes, and since 5.2 indicates that forecasting difficulty varies across these calendars, we examine whether the order of interpretability is stable by market. For each metric, we rank the eleven approved arms by mean value in each market, then correlate the rankings across market pairs (see Table 13). Agreement is consistently high: across the three metrics and three market pairs, the nine rank correlations range from 0.88 to 0.99. None of the conclusions stated above are based on peculiarities of any one market. Together with the

findings in 5.2, this shows that forecast difficulty depends on the market, while the ordering of the model-explainer pairs does not.

We carried out a complete-block analysis. The matched-pair comparisons mentioned above include all stocks on which a particular pair of arms was admitted, and this collection varies from one pair to another. To check the robustness of the results, we also applied the Friedman test with Nemenyi post hoc comparisons, limiting the analysis to stocks for which all arms in each group were admitted simultaneously. Fig. 8 shows the mean ranks alongside a single critical difference; Fig. 9 presents the same pairwise test for the price-lag panel. The two analyses are algebraically consistent in that a pair is separated in Fig. 9 if and only if its mean ranks differ by more than the width of the band in (a). The order matches the matched-pair analysis, with arms outside the critical-difference band separable from the best-ranked arm and pairs with $p < 0.05$ separating in pairwise comparisons. We treat this as a secondary analysis because the complete-block condition is very strict in the context of a skill gate: only 11 stocks allow all arms to be tested on the price-lag panel, whereas each matched-pair comparison has 10 to 18 stocks. This stock scarcity, not any discrepancy between the two analyses, is why the matched-pair tests take priority throughout.

Table 13 Spearman correlation of mean arm ranks between markets

Metric	Europe – Japan	Europe – US	Japan – US
Temporal agreement (ρ)	0.964	0.891	0.882
Temporal rank-biased overlap	0.955	0.991	0.936
Input stability	0.991	0.982	0.964

Multiplicity. Of 250 pairwise arm comparisons across all metrics and feature sets, 148 reach Holm-adjusted $p < 0.10$ within family, and we also report a second Holm pass across all families. As noted in 5.4, roughly a third of the significant results —

approximately 51 of 148 — belong to the AOPC family and should not be read as faithfulness findings.

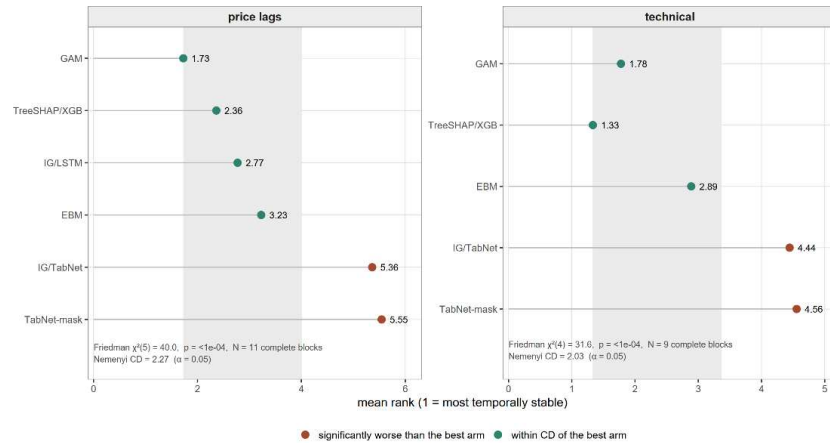


Fig. 8 Friedman test with Nemenyi confirmation

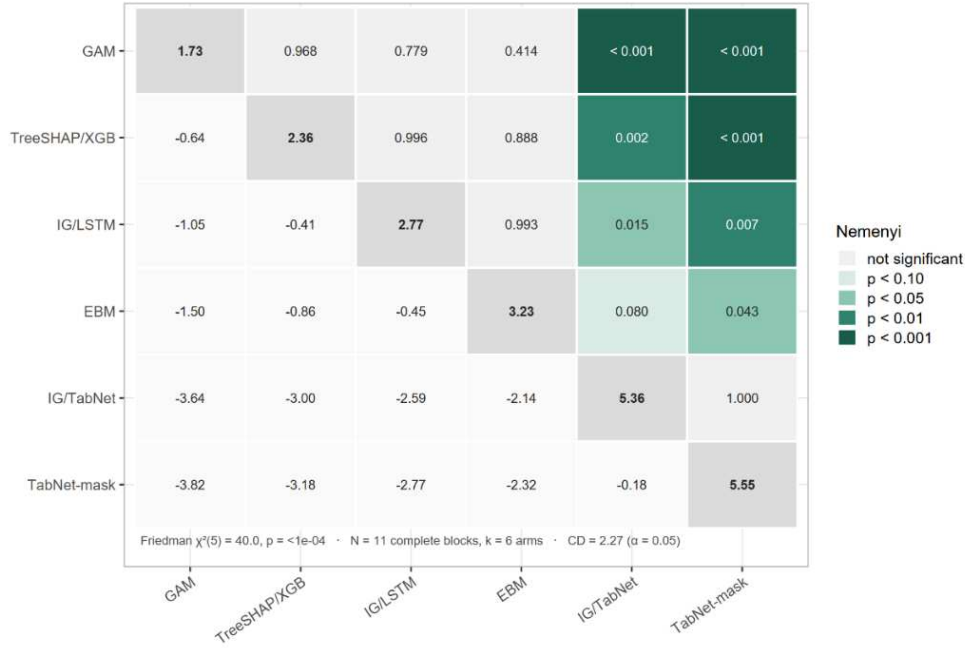


Fig. 9 Nemenyi matrix

6 Discussion

6.1 The paradigm does not determine stability, but the pairing does

This study aimed to examine the relative reliability of intrinsic and post hoc explanations. The results do not provide a clear answer based just on the distinction between these two types of explanations. Instead, they show that reliability depends on the model–representation–explainer configuration assessed under a given perturbation. If either qualifier is omitted, the conclusion changes.

The study's design enables a compelling matched two-by-two comparison. If the explainer is held constant and the model architecture is changed, the Integrated Gradients values for the LSTM and TabNet differ by a median of 0.519 in temporal rank agreement ($n = 11$, $p = 0.014$) and by 0.076 in input stability ($p = 0.006$). In contrast, when the architecture is held constant and the explainer is varied, the intrinsic mask and Integrated Gradients, both used with the same TabNet network, are statistically indistinguishable in temporal agreement across all measures ($p > 0.10$). Notably, the same attribution method ends up at both ends of the price-lag stability ordering: IG/LSTM ranks second in input stability, while IG/TabNet ranks last. By contrast, two different methods on the same network are indistinguishable. The TabNet arms have temporal agreement ranging from 0.054 to 0.117, while all other arms fall between 0.43 and 0.64; furthermore, each comparison between

TabNet and the other arms passes the Holm correction for all four agreement measures in both feature sets.

These results support the argument put forward by Madsen et al. [38]. They claim that the distinction between intrinsic and post-hoc approaches is not the most useful one, and they offer similar evidence for that view. The results also confirm, rather than contradict, Rudin's [26] recommendation. Rudin's criticism of post-hoc explanations for black-box models rests on the concern that such approximations might misrepresent models they did not help construct. The current findings show no sign of this kind of failure when assessed using temporal or ablation methods: the post-hoc attribution assigned to a TabNet network is as good as the network's own intrinsic explanation. Yet the results show that an intrinsic explanation can be unreliable in ways that go beyond what Rudin's argument anticipated.

One model, the spline GAM based on price lags, conforms to Rudin's recommendations as measured by its own standards. It comes in first or is jointly first when assessed using three separate criteria: its accuracy is statistically no different from that of the HAR-RV benchmark and it outperforms all of the black-box models in the study; it attains the highest level of temporal agreement; and it produces identical results across different random seeds. As a result, the additive model, which is designed to be interpretable, shows no measurable decline in either accuracy or stability. TabNet is noteworthy because, although it is also described as interpretable by design, it finishes last in both accuracy relative to the benchmark and explanation reproducibility. Native interpretability is not a simple yes-or-no feature; in the additive model,

the explanation directly mirrors its functional form, while the attention mask results from an optimization process that the data largely leave unconstrained.

In each of the three markets, both parts of this contrast stand on their own. When we rank the eleven admitted arms by temporal agreement within each market and then compare the rankings across market pairs, the resulting values range from 0.88 to 0.99, and the TabNet arms are at or below zero in all three cases. In contrast, all other arms fall between 0.39 and 0.69 (Section 5.7). The conclusion that the model dominates the explainer is therefore not an artifact resulting from pooling: it appears three times on samples that have different trading calendars, different liquidity profiles, and different volatility regimes, and in two of these markets the intrinsic mask's temporal agreement is actually negative, which means that its importance ordering reverses between consecutive quarters.

6.2 A dissociation: the more stable of two explanations will depend on the perturbation

In one trained network examined by two explainers, the explainers' order reverses when perturbations are applied. With input jitter, the intrinsic mask is more stable and achieves a higher rank for every matched

stock in both feature representations, with a rank-biserial correlation of exactly +1.000. However, when we re-estimate the mask with new random seeds, it becomes considerably less stable, with a rank-biserial correlation ranging from -0.51 to -0.93 and the same direction across all three agreement measures. In the case of temporal shift and under ablation, the two results are indistinguishable. Since both sets of explainers use the same network—a fact we verified rather than assumed through a state-invariance check on each attribution call—no difference in the underlying model can explain the reversal.

The reversal also rules out the most straightforward objection to the input-stability result when considered by itself. Since Integrated Gradients is gradient-based, it is, by design, sensitive to input perturbations, while an attention mask is selection-based; this definitional advantage would mean the mask should perform better under all perturbations. It does not. The two measures capture genuinely different properties, and the mask performs better on one and worse on the other.

The same stability holds for the input-stability aspect of the reversal: when the arms are ranked by input stability in each market, rank correlations between markets range from 0.96 to 0.99, the highest of the three correlation types examined in that section (Section 5.7). The dissociation is thus not a characteristic of the data from one particular market.

However, the design does not rule out the opposite limitation mentioned in Section 6.7: it is based on a single architecture.

The proposal is a hypothesis rather than a firm conclusion. TabNet's sparse attention mechanism selects a subset of features at each stage of the decision process. For a given parameterization, this selection seems to change smoothly in response to small input changes, which could account for the mask's robustness to input jitter. However, the features the network chooses are determined by the optimization process, which the data does not strongly constrain. The seed results indicate that the rank agreement between independently seeded networks is 0.006 for price lags and 0.086 for technical indicators, while the predictions from those networks correlate at 0.27 and / 0.25, respectively. On the other hand, Integrated Gradients collects gradients of the learned function along a path from a baseline, which may introduce pointwise noise but reduces reliance on the optimizer's selection pattern. To test this hypothesis, it would be necessary to break down the seed variance of the mask by attention step, a task which is left for future research.

In discussing attention as an explanation, the present findings support a skeptical view that has not been recorded in this area. Jain and Wallace [63] showed that different attention distributions can lead to the same predictions, whereas Wiegrefe and Pinter [64] maintained that this does not necessarily rule out attention as an explanation since it depends on the claim one is making. Seed reproducibility constitutes a separate and perhaps more rigorous test; the rank agreement of 0.047 among networks with similar predictions shows that the architecture's assignment of importance to the input is almost entirely arbitrary with respect to the training run. Therefore, the feature that underpins TabNet's claim of being interpretable by design is the least reproducible of all the explanations examined.

6.3 Explanation stability does not serve as evidence of faithfulness

Stability is generally seen as a desirable feature of explanations because an explanation that changes with minor perturbations cannot indicate a lasting relationship. Alvarez-Melis and Jaakkola [59] promoted stability as an explicit evaluation criterion, and it appears in almost all benchmarks discussed in Section 2.3. Yet the seed results from this study show that the inference from stability to trustworthiness does not usually hold.

The explanations produced by both LSTM models are more reproducible across different seeds than their predictions. This is clear in the technical-indicator

arm: predictions from independently seeded networks correlate at 0.190, whereas their Integrated Gradients attributions agree at 0.493. Although this arm fails to exceed the skill gate for any of the 20 stocks and its learned function shows only slight run-to-run reproducibility, the attributions themselves are consistent. However, this consistency does not mean it captures a true relationship, since it reproduces hardly any substantial pattern. A more reasonable explanation is that the attributions reflect the geometry of the input representation, particularly the correlation structure described in Section 5.3, rather than the networks' learning process. Raising the epoch budget lowered the price-lag gap from +0.266 to +0.095 but did not eliminate it.

Another result supports this conclusion from a different angle. When the EBM is applied to the price-lag representation, it reliably reproduces its predictions across different seeds (coefficient of variation: 0.014), whereas its explanations show poor reproducibility (rank agreement: 0.458) and only moderate temporal stability. The example therefore illustrates a situation in which a model may be stable over time but fragile when the estimates are recomputed. Temporal stability and seed stability are separate properties, and reporting a single stability measure obscures these differences.

This has important consequences for how we evaluate models. Stability measures based on a single trained model, this being the case for most of the metrics mentioned in Section 2.3, are fundamentally incapable of picking up on this kind of failure. To detect it, you must re-estimate the model and then compare the agreement of the predictions with the agreement of the explanations, a procedure that is considerably less costly than the original training and therefore should be part of any evaluation aimed at supporting reliance on an explanation. The link with the multiplicity literature is obvious: if the training pipeline does not specify clearly which features the model depends on, then an explanation obtained from a single run refers only to one member of a Rashomon set rather than to a property of the data, and the stability of the explanation under input perturbation does not indicate which is the case.

6.4 What the faithfulness null shows and does not show

This task provides no positive evidence that ablation-based faithfulness distinguishes the explanation rankings from a random ordering. For ten of the eleven admitted arms, the difference between importance-ordered AOPC and the corresponding random-order control is close to zero, ranging from -0.00014 to $+0.00012$. The only substantial exception is the GAM with technical indicators, for which

importance-ordered ablation produces an AOPC of 0.0055 compared with 0.0119 under random-order ablation, yielding a difference of -0.0064 . Importantly, this exception runs opposite to what a faithful ranking would predict: random ablation degrades predictive performance more strongly than importance-ordered ablation. Thus, none of the eleven admitted arms provides convincing evidence that the explanation-based feature ordering is more faithful than a random ordering.

The measured correlation structure indicates that collinearity can explain only part of the weak faithfulness results. For the technical-indicator panel, this explanation is plausible because 23% of feature pairs have $|r| > 0.8$, while 13 nominal features span only about 4.3 effective dimensions. By contrast, the price-lag panel is essentially orthogonal, with a mean absolute pairwise correlation of 0.029, a condition number of 1.13, and an effective rank of 5.98 out of 6, yet the faithfulness metric remains similarly uninformative. Two mechanisms therefore appear to operate: feature substitutability in the technical-indicator panel and, in the price-lag panel, a signal-to-noise ratio so low that removing any single feature changes out-of-sample error by an amount indistinguishable from sampling variation. Of these two mechanisms, only the low signal-to-noise explanation can account for the weak faithfulness results across both feature representations.

The result suggests a task-metric interaction: under the low signal-to-noise conditions examined here, AOPC does not distinguish explanation-based rankings from random ordering. Bragança and Souto [57] find that removing the top decile of SHAP-ranked features collapses classification performance—that is, extreme faithfulness by this criterion—when applied to time-series classification tasks. The fact that these two findings are not in contradiction lies in the fact that ablation faithfulness assumes that removing an important feature results in a measurable deterioration of performance, which in turn requires a usable signal-to-noise ratio; a five-day realized volatility forecast based on daily equity data does not provide such a ratio, whereas a classification task with separable classes does. This finding is the boundary condition, and it implies that when the target is noisy, ablation-based faithfulness should be reported alongside a randomized control rather than assumed to transfer.

This directly affects how to interpret the pairwise results in Section 5. About one third of the arm comparisons that remain after applying the Holm correction are part of the AOPC family and have effect sizes of the order of 10^{-4} to 10^{-3} on a metric that does not outperform a random-order baseline. These

represent statistically significant differences in a quantity that contains no information, and they show a general danger: while multiplicity correction controls the false-positive rate, it says nothing about whether the underlying measurement is meaningful. A randomized controlled trial involving the metric itself is the kind of check that multiplicity correction cannot offer.

6.5 Problems with measurement in explanation agreement

The range of a fixed-depth top- k overlap measure depends on the feature space size. Any two five-element subsets of a six-element set share at least four elements, so on the price-lag panel the top-5 Jaccard index can take only two values, 0.667 and 1.000. On the thirteen-feature panel, the same index has a floor of 0.000 and six attainable values, drawn from the $C(13,5) = 1,287$ possible five-feature subsets. Its expected value under independent random rankings is 0.7222 in the first case and 0.2537 in the second. A direct comparison of 0.854 with 0.549 therefore compares two quantities on different scales, and the difference it reports is largely the difference in their floors.

The figures for each individual stock show just how serious the problem is in practice. For XOM and 6758.T, the two explainers rank the six price-lag features in opposite orders, with Spearman correlations of -0.143 and -0.257 , respectively; in both cases, the top-5 Jaccard score is 0.667, the highest possible level of disagreement. For 7203.T, the measure is 1.000—indicating complete agreement—while the Spearman correlation is 0.143. After eliminating the random-ranking floor, seven of the sixteen admitted price-lag cells fall below chance agreement, even though the raw measure is reassuringly 0.667. Of the four size-invariant or floor-corrected measures, three measures that we report, all of which are size-invariant or have been explicitly floor-corrected, show higher agreement on the thirteen-feature panel. The only exception is the floor-corrected fixed- k Jaccard, which still gives the advantage to price lags since the median value for that panel is at the measure's ceiling, so that the correction has nothing to work with — this is a kind of saturation effect that is different from the floor problem and one that the fixed-depth overlap method cannot avoid when the number of features is small.

To interpret this meaningfully, one must go beyond dimensionality. The correlation structure obtained shows that the thirteen-feature panel covers about 4.3 effective dimensions, whereas the six-feature panel covers 5.98 effective dimensions. Thus, the two groups show an inverse relationship between nominal width and effective dimensionality. Agreement

among the explainers is greater for the panel, which is nominally wider but effectively narrower, suggesting that the results reflect the influence of effective dimensionality and correlation structure. This reversal was not foreseen in the study design; we discovered it through diagnostic analysis, and it underpins the interpretability of the corrected result.

We offer it here as a methodological contribution rather than merely a correction of earlier work. Top- k overlap measures are frequently used to assess agreement between explanations, especially in scenarios with varying numbers of features, for example, across different datasets, under different feature-selection procedures, or before and after dimensionality reduction. In such comparisons, the measure's floor varies with the setting. You can remove this confounding effect at very little cost by using rank-biased overlap [65] or by explicitly normalizing against the expectation under random ranking. Therefore, adopt one of these methods as standard practice when comparing agreement across representations of different widths.

6.6 Practical Implications for explainer Selection

In practice, this dissociation means you should consider the distinction between intrinsic and post hoc explanations depending on the application. Explanations designed to support individual local decisions must be robust to measurement noise, while those intended to support global importance rankings must remain consistent as the model is updated; they face different kinds of perturbations and should therefore be chosen accordingly.

6.7 Limitations and Future Directions

Several limitations should be considered when interpreting these findings. First, the same-model comparison between intrinsic and post-hoc explanations is based only on TabNet, the only architecture in this study that provides both an intrinsic attention mask and a post-hoc Integrated Gradients (IG) explanation. Consequently, the divergence reported in Section 6.2 cannot establish whether the observed pattern reflects a general distinction between intrinsic and post-hoc explanations or a property specific to sparse attention. Moreover, the two explainers quantify different concepts: IG attributes prediction changes relative to a baseline, whereas the TabNet mask represents feature-selection magnitude without a baseline. Some disagreement is therefore inherent in their definitions and should not be interpreted as evidence for or against either explanation paradigm. Future work could address this limitation by applying post-hoc kernel-based Shapley attribution to the same fitted GAM and EBM models, creating three same-model

comparisons without requiring new data or a different evaluation protocol. Decomposing TabNet mask variability by decision step could also clarify the mechanism underlying the perturbation reversal observed in Section 6.2.

The stability results also require careful interpretation. Low temporal stability may indicate either genuine adaptation to changing market regimes or explanation noise, and the present design cannot distinguish between them. Seed stability provides a cleaner measure of re-estimation uncertainty because the underlying data remain unchanged; however, GAM is deterministic under iteratively reweighted least squares and therefore produces perfect seed stability by construction. This serves as a useful validation check but prevents direct comparison with stochastic learners on this dimension. Several design choices may also affect the results, including the 8% feature-level noise used for input perturbation, the five seeds applied to the final three folds, and the $\text{RelMAE} < 1$ skill gate. These parameters were not tuned, although Section 5.7 reports a stricter sensitivity criterion requiring a Benjamini–Hochberg-adjusted $q \leq 0.10$.

External validity is another limitation. The study examines 20 large-capitalization equities, one forecast horizon, one volatility target, and a restricted set of feature representations. All explanations are correlational and should not be interpreted causally. Fold availability also differs across feature panels and regions, producing unequal evidence across cells. Predictive skill varies considerably by market: among eligible cells, median RelMAE is 0.870 in Japan and 0.882 in Europe, compared with 1.001 in the United States, where step-(h) persistence is more competitive. Consequently, the accuracy results should not be assumed to transfer directly across markets, even though Section 5.7 suggests that the interpretability ranking is more stable. Because fewer US cells pass the skill gate, the admitted sample becomes relatively balanced across regions (59 US, 52 European, and 54 Japanese cells), rather than reflecting the original 10–5–5 stock distribution. This changes the weighting of pooled interpretability results, although the observed cross-market rank agreement reduces concern about substantial distortion.

Future research should extend the framework in several directions. Adding absolute returns, squared returns, and realized-variance lags could determine how much of HAR-RV's accuracy advantage results from its explicit volatility-magnitude information, although this would reduce the intended near-orthogonality of the price-lag panel. More importantly, the three-axis perturbation framework—

input, temporal, and re-estimation stability—should be tested across other asset classes, forecast horizons, and financial targets to determine whether the weak faithfulness results are specific to equity-volatility forecasting. This perturbation taxonomy may also serve as a broader XAI evaluation protocol: relying on a single stability measure could lead to opposite conclusions depending solely on which perturbation is chosen, whereas separating the three sources of instability provides a more informative assessment of explanation robustness.

7 Conclusion

The study investigated whether intrinsic or post hoc explanations are the more reliable option when making machine-learning forecasts of volatility. To examine this question under conditions where the answer should hold, it used purged walk-forward validation on 20 equities from three regions, applied a skill gate that allowed only models that outperformed a persistence baseline out of sample, included the HAR-RV benchmarks, and carried out matched-stock paired inference throughout.

No answer to the question is independent of the perturbation level. With one trained TabNet network, the order in which the two explanations are presented reverses depending on what they must endure: the intrinsic attention mask proves more resistant to input perturbations for each stock examined and is considerably less resistant when the model is re-estimated using new seeds. Across the benchmark, explanation stability depends mainly on which model is explained rather than on the explainer's family—the same attribution method ends up at both ends of the stability spectrum when applied to the two architectures. By contrast, two distinct methods analyzing the same network yield indistinguishable results.

Two findings concern how to evaluate explanations, not which to choose. Explanation stability is not evidence of faithfulness: our weakest arm produced attributions more reproducible across seeds than the predictions those attributions described, a failure mode invisible to any evaluation performed on a single trained model. And ablation-based faithfulness carried no signal on this task, performing no better than random feature ordering for every arm—a boundary condition on a widely used metric that only a randomized control makes visible. A separate measurement artifact reversed one of our own earlier findings: the apparent erosion of explainer agreement with feature count is a property of fixed-depth top-k overlap, whose floor rises as the feature space narrows, and size-invariant measures place agreement higher on the wider representation.

The key implication is that explanation reliability is not inherent to a particular paradigm. Instead, it is a property of a model–representation–explainer configuration, evaluated under a specific perturbation, and must be empirically measured rather than inferred from the architecture's intended design. In this study, the architecture marketed as interpretable-by-design produced the least reproducible explanation across all arms. For practitioners, this suggests that the choice of explainer should depend on the use case: intrinsic explanations are preferable when local attributions must withstand input noise, while post-hoc attributions are more suitable when global importance must persist through model refreshes. In all cases, a re-estimation check should precede reliance on an explanation for audit or risk-related decisions.

Two constraints bound these conclusions. The same-model contrast relies on a single architecture, so whether the perturbation reversal generalizes beyond sparse attention remains open; replicating it on additive models with post hoc Shapley attribution is the natural next step and requires no new data. And a three-parameter HAR-RV regression achieved the best median accuracy in the study, decisively beating the neural and tree-based arms while remaining statistically indistinguishable from the additive one. We report this as a constraint on the forecasting claims rather than on the interpretability findings: we selected the models whose explanations we compare for demonstrable out-of-sample skill, not for being the best available forecasters.

Statements and Declarations

Funding

We received no funding for this study.

Competing Interests

The authors have no relevant financial or non-financial interests to disclose.

Author Contributions

Conceptualization: [HD]; Methodology: [HD]; Software: [HD]; Validation: [AY]; Formal analysis: [SND]; Visualization: [HD]; Writing—original draft preparation: [HD]; Writing—review and editing: [AY]; Supervision: [AND]. All authors read and approved the final manuscript.

Data Availability

We obtained the market data used in this study from Yahoo Finance, a publicly accessible source of historical financial market data. The dataset

comprises daily adjusted closing prices, high and low prices, and trading volumes for the 20 equities included in the study from 2016–2025. We derived the variables used for model development from these raw market data using the feature-engineering procedures described in the manuscript. The processed datasets supporting this study's findings are available from the corresponding author upon reasonable request.

Code Availability

The code used for data preprocessing, model estimation, explainability analysis, statistical testing, and generation of the reported results is available from the corresponding author upon reasonable request.

Ethics Approval and Consent to Participate

Not applicable. This study uses publicly available financial market data and does not involve human participants, animals, personal data, or biological materials.

Consent for Publication

Not applicable.

References

1. Christensen K, Siggaard M, Veliyev B (2023) A Machine Learning Approach to Volatility Forecasting. *J Financ Econom* 21:1680–1727. <https://doi.org/10.1093/jfinec/nbac020>
2. Zhang C, Zhang Y, Cucuringu M, Qian Z (2024) Volatility Forecasting with Machine Learning and Intraday Commonality. *J Financ Econom* 22:492–530. <https://doi.org/10.1093/jfinec/nbad005>
3. Arratia A, El Daou M, Kagerhuber J, Smolyarova Y (2025) Examining Challenges in Implied Volatility Forecasting: A Critical Review of Data Leakage and Feature Engineering combined with High-Complexity Models. *Comput Econ*. <https://doi.org/10.1007/s10614-025-11172-z>
4. Bucci A (2020) Realized Volatility Forecasting with Neural Networks. *J Financ Econom* 18:502–531. <https://doi.org/10.1093/jfinec/nbaa008>
5. Leushuis RM, Petkov N (2026) Advances in forecasting realized volatility: a review of methodologies. *Financ Innov* 12:14. <https://doi.org/10.1186/s40854-025-00809-5>

6. Rigamonti A (2024) Can machine learning make technical analysis work? *Financ Mark Portf Manag* 38:399–412. <https://doi.org/10.1007/s11408-024-00451-8>
7. Saud AS, Shakya S (2024) Technical indicator empowered intelligent strategies to predict stock trading signals. *J Open Innov Technol Mark Complex* 10:100398. <https://doi.org/10.1016/j.joitmc.2024.100398>
8. Mostafavi SM, Hooman AR (2025) Key technical indicators for stock market prediction. *Mach Learn Appl* 20:100631. <https://doi.org/10.1016/j.mlwa.2025.100631>
9. Raza H, Akhtar Z (2024) Predicting stock prices in the Pakistan market using machine learning and technical indicators. *Mod Finance* 2:46–63. <https://doi.org/10.61351/mf.v2i2.167>
10. Goswami B, Uddin A (2025) Significance of predictors: revisiting stock return predictions using explainable AI. *Ann Oper Res*. <https://doi.org/10.1007/s10479-025-06717-2>
11. Bagnara M (2024) Asset Pricing and Machine Learning: A critical review. *J Econ Surv* 38:27–56. <https://doi.org/10.1111/joes.12532>
12. Gunnarsson ES, Isern HR, Kaloudis A, Risstad M, Vigdel B, Westgaard S (2024) Prediction of realized volatility and implied volatility indices using AI and machine learning: A review. *Int Rev Financ Anal* 93:103221. <https://doi.org/10.1016/j.irfa.2024.103221>
13. Massimiliano F, Tiziana C (2024) Improving the Interpretability of Asset Pricing Models by Explainable AI: A Machine Learning-based Approach. *Econ Comput Econ Cybern Stud Res* 58:5–19. <https://doi.org/10.24818/18423264/58.4.24.01>
14. Hastie T, Tibshirani R (1986) Generalized Additive Models. *Stat Sci* 1:. <https://doi.org/10.1214/ss/1177013604>
15. Lou Y, Caruana R, Gehrke J, Hooker G (2013) Accurate intelligible models with pairwise interactions. In: *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, Chicago Illinois USA, pp 623–631
16. Arik SÖ, Pfister T (2021) TabNet: Attentive Interpretable Tabular Learning. *Proc AAAI Conf Artif Intell* 35:6679–6687. <https://doi.org/10.1609/aaai.v35i8.16826>
17. Lundberg S, Lee S-I (2017) A Unified Approach to Interpreting Model Predictions
18. Lundberg SM, Erion G, Chen H, DeGrave A, Prutkin JM, Nair B, Katz R, Himmelfarb J, Bansal N, Lee S-I (2019) Explainable AI for Trees: From Local Explanations to Global Understanding
19. Wachter S, Mittelstadt B, Russell C (2017) Counterfactual Explanations Without Opening the Black Box: Automated Decisions and the GDPR. *SSRN Electron J*. <https://doi.org/10.2139/ssrn.3063289>
20. Sundararajan M, Taly A, Yan Q (2017) Axiomatic Attribution for Deep Networks
21. Guidotti R (2024) Counterfactual explanations and how to find them: literature review and benchmarking. *Data Min Knowl Discov* 38:2770–2824. <https://doi.org/10.1007/s10618-022-00831-6>
22. Ranjan P, Itani R, Faccia A (2025) An Interpretable 1D-CNN Framework for Stock Price Forecasting: A Comparative Study with LSTM and ARIMA. *FinTech* 4:63. <https://doi.org/10.3390/fintech4040063>
23. Yeo WJ, Van Der Heever W, Mao R, Cambria E, Satapathy R, Mengaldo G (2025) A comprehensive review on financial explainable AI. *Artif Intell Rev* 58:189. <https://doi.org/10.1007/s10462-024-11077-7>
24. Theissler A, Spinnato F, Schlegel U, Guidotti R (2022) Explainable AI for Time Series Classification: A Review, Taxonomy and Research Directions. *IEEE Access* 10:100700–100724. <https://doi.org/10.1109/ACCESS.2022.3207765>
25. Khan FS, Mazhar SS, Mazhar K, A. AlSaleh D, Mazhar A (2025) Model-agnostic explainable artificial intelligence methods in finance: a systematic review, recent developments, limitations, challenges and future directions. *Artif Intell Rev* 58:232. <https://doi.org/10.1007/s10462-025-11215-9>
26. Rudin C (2019) Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nat Mach Intell* 1:206–215. <https://doi.org/10.1038/s42256-019-0048-x>
27. Laugel T, Lesot M-J, Marsala C, Renard X, Detryniecki M (2019) The Dangers of Post-hoc Interpretability: Unjustified Counterfactual Explanations. In: *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence*. International Joint

- Conferences on Artificial Intelligence Organization, Macao, China, pp 2801–2807
28. Wiratsin I-O, Ragkhitwetsagul C (2025) Effectiveness of Explainable Artificial Intelligence (XAI) Techniques for Improving Human Trust in Machine Learning Models: A Systematic Literature Review. *IEEE Access* 13:121326–121350. <https://doi.org/10.1109/ACCESS.2025.3575022>
 29. Haque AB, Islam AKMN, Mikalef P (2023) Explainable Artificial Intelligence (XAI) from a user perspective: A synthesis of prior literature and problematizing avenues for future research. *Technol Forecast Soc Change* 186:122120. <https://doi.org/10.1016/j.techfore.2022.122120>
 30. Jacovi A, Goldberg Y (2020) Towards Faithfully Interpretable NLP Systems: How should we define and evaluate faithfulness?
 31. Bilodeau B, Jaques N, Koh PW, Kim B (2024) Impossibility theorems for feature attribution. *Proc Natl Acad Sci* 121:e2304406120. <https://doi.org/10.1073/pnas.2304406120>
 32. Arsenault P-D, Wang S, Patenaude J-M (2025) A Survey of Explainable Artificial Intelligence (XAI) in Financial Time Series Forecasting. *ACM Comput Surv* 57:1–37. <https://doi.org/10.1145/3729531>
 33. Černevičienė J, Kabašinskas A (2024) Explainable artificial intelligence (XAI) in finance: a systematic literature review. *Artif Intell Rev* 57:216. <https://doi.org/10.1007/s10462-024-10854-8>
 34. Mohamed A, Abdelqader K, Shaalan K (2025) Explainable Artificial Intelligence: A systematic Review of Progress and Challenges. *Intell Syst Appl* 28:200595. <https://doi.org/10.1016/j.iswa.2025.200595>
 35. Sharbati A, Noorollahi Y, Hajinezhad A, Yousefi H (2026) Decoding the energy paradox: An interpretable XGBoost-SHAP framework for forecasting decoupled demand futures. *Results Eng* 29:109741. <https://doi.org/10.1016/j.rineng.2026.109741>
 36. Suprihadi E, Danila N, Ali Z, Ananta GP (2026) Explainable Ensemble Learning for Predicting Stock Market Crises: Calibration, Threshold Optimization, and Robustness Analysis. *Information* 17:114. <https://doi.org/10.3390/info17020114>
 37. Freeborough W, Van Zyl T (2022) Investigating Explainability Methods in Recurrent Neural Network Architectures for Financial Time Series Data. *Appl Sci* 12:1427. <https://doi.org/10.3390/app12031427>
 38. Madsen A, Lakkaraju H, Reddy S, Chandar S (2024) Interpretability Needs a New Paradigm
 39. Zschech P, Weinzierl S, Kraus M (2026) Inherently Interpretable Machine Learning: A Contrasting Paradigm to Post-hoc Explainable AI. *Bus Inf Syst Eng* 68:445–463. <https://doi.org/10.1007/s12599-025-00964-0>
 40. Shu J, Yu H, Liu G, Yang H, Guo W, Phoon C, Alfred S, Hu H (2024) Proposing an inherently interpretable machine learning model for shear strength prediction of reinforced concrete beams with stirrups. *Case Stud Constr Mater* 20:e03350. <https://doi.org/10.1016/j.cscm.2024.e03350>
 41. Brocki L, Chung NC (2019) Concept Saliency Maps to Visualize Relevant Features in Deep Generative Models
 42. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D (2017) Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. In: 2017 IEEE International Conference on Computer Vision (ICCV). IEEE, Venice, pp 618–626
 43. Štureka D, Lazarova-Molnar S (2025) Surrogate Modeling: Review and Opportunities for Expert Knowledge Integration. *Procedia Comput Sci* 257:826–833. <https://doi.org/10.1016/j.procs.2025.03.106>
 44. Fu Z, Yang X, Ma Y, Sun Y, Wang T (2025) Integrating explainable AI and causal inference to unveil regional air quality drivers in China. *J Environ Manage* 390:126270. <https://doi.org/10.1016/j.jenvman.2025.126270>
 45. Munoz C, da Costa K, Modenesi B, Koshiyama A (2023) Evaluating Explainability in Machine Learning Predictions through Explainer-Agnostic Metrics
 46. Mersha M, Lam K, Wood J, AlShami AK, Kalita J (2024) Explainable artificial intelligence: A survey of needs, techniques, applications, and future direction. *Neurocomputing* 599:128111. <https://doi.org/10.1016/j.neucom.2024.128111>
 47. Ai Q, Narayanan R L (2021) Model-agnostic vs. Model-intrinsic Interpretability for Explainable Product Search. In: Proceedings of the 30th ACM International Conference on Information & Knowledge Management. ACM, Virtual Event Queensland Australia, pp 5–15

48. Puthanveettil Madathil A, Luo X, Liu Q, Walker C, Madarkar R, Cai Y, Liu Z, Chang W, Qin Y (2024) Intrinsic and post-hoc XAI approaches for fingerprint identification and response prediction in smart manufacturing processes. *J Intell Manuf* 35:4159–4180. <https://doi.org/10.1007/s10845-023-02266-2>
49. Chakraborty A, Karagoz G, Meratnia N (2026) A Cross-Domain Benchmark of Intrinsic and Post Hoc Explainability for 3D Deep Learning Models. *J Imaging* 12:63. <https://doi.org/10.3390/jimaging12020063>
50. Ammar OH, Rejeb A, Rejeb K (2026) Explainable artificial intelligence in finance: A bibliometric review. *Finance Res Open* 2:100131. <https://doi.org/10.1016/j.finr.2026.100131>
51. Sailer L (2026) Explainable artificial intelligence in finance: Mapping business problems and technological solutions. *Finance Res Open* 2:100128. <https://doi.org/10.1016/j.finr.2026.100128>
52. Metsch JM, Hauschild A-C (2025) BenchXAI: Comprehensive benchmarking of post-hoc explainable AI methods on multi-modal biomedical data. *Comput Biol Med* 191:110124. <https://doi.org/10.1016/j.compbiomed.2025.110124>
53. Bodria F, Giannotti F, Guidotti R, Naretto F, Pedreschi D, Rinzivillo S (2023) Benchmarking and survey of explanation methods for black box models. *Data Min Knowl Discov* 37:1719–1778. <https://doi.org/10.1007/s10618-023-00933-9>
54. Belaid MK, Bornemann R, Rabus M, Krestel R, Hüllermeier E (2023) Compare-xAI: Toward Unifying Functional Testing Methods for Post-hoc XAI Algorithms into a Multi-dimensional Benchmark. In: Longo L (ed) *Explainable Artificial Intelligence*. Springer Nature Switzerland, Cham, pp 88–109
55. Bragança H, Souto E (2026) Unified time series classification framework for explainable artificial intelligence. *Sci Rep* 16:22148. <https://doi.org/10.1038/s41598-026-49467-2>
56. Samek W, Binder A, Montavon G, Lapuschkin S, Müller K-R (2017) Evaluating the Visualization of What a Deep Neural Network Has Learned. *IEEE Trans Neural Netw Learn Syst* 28:2660–2673. <https://doi.org/10.1109/TNNLS.2016.2599820>
57. Alvarez-Melis D, Jaakkola TS (2018) On the Robustness of Interpretability Methods
58. Krishna S, Han T, Gu A, Wu S, Jabbari S, Lakkaraju H (2022) The Disagreement Problem in Explainable Machine Learning: A Practitioner’s Perspective
59. Brughmans D, Melis L, Martens D (2024) Disagreement amongst counterfactual explanations: how transparency can be misleading. *TOP* 32:429–462. <https://doi.org/10.1007/s11750-024-00670-2>
60. Krauss C, Do XA, Huck N (2017) Deep neural networks, gradient-boosted trees, random forests: Statistical arbitrage on the S&P 500. *Eur J Oper Res* 259:689–702. <https://doi.org/10.1016/j.ejor.2016.10.031>
61. Fischer T, Krauss C (2018) Deep learning with long short-term memory networks for financial market predictions. *Eur J Oper Res* 270:654–669. <https://doi.org/10.1016/j.ejor.2017.11.054>
62. Gu S, Kelly B, Xiu D (2020) Empirical Asset Pricing via Machine Learning. *Rev Financ Stud* 33:2223–2273. <https://doi.org/10.1093/rfs/hhaa009>
63. Jain S, Wallace BC (2019) Attention is not Explanation. In: *Proceedings of the 2019 Conference of the North. Association for Computational Linguistics*, Minneapolis, Minnesota, pp 3543–3556
64. Wiegrefe S, Pinter Y (2019) Attention is not not Explanation. In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Association for Computational Linguistics, Hong Kong, China, pp 11–20

