



iJRASET

International Journal For Research in
Applied Science and Engineering Technology



INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 14 **Issue:** X **Month of publication:** October 2026

DOI: <https://doi.org/10.22214/ijraset.2026.85252>

www.ijraset.com

Call: ☎ 08813907089

E-mail ID: ijraset@gmail.com

Explainable Artificial Intelligence for Stock Market Trend Prediction and Risk Assessment: A Review

Pallavi Pathare¹, Bhagyashri Sanap², Pooja Bhagwat³, Sanika Wagh⁴, Shantanu Nande⁵

Department of Computer Engineering, Sir Visvesvaraya Institute of Technology (SVIT), Nashik, Savitribai Phule Pune University, Maharashtra, India

Abstract: Machine learning models are now widely used to forecast stock price movements and to measure financial risk, yet most cannot explain their own outputs. In finance that is a serious weakness, because forecasts influence real investment decisions and regulators expect firms to justify the models they use. This review asks how explainable artificial intelligence (XAI) has been applied to stock trend prediction and risk assessment, and how far the resulting explanations can be trusted. We first outline the economic theory of predictability and the standard risk measures, then set out the mathematics of the main explanation methods: Shapley-value attribution, LIME, gradient-based attribution, attention, and counterfactuals. We then compare representative studies, summarise regulatory expectations, and propose a five-layer framework for explainable forecasting. Across the literature we find that explanations are most useful as a validation tool, and that they become un-reliable when features are correlated, when markets change regime, or when a model is overfitted. Evidence that explanations improve decisions remains thin. We close with open problems, including causal explanation, shared benchmarks, and monitoring of explanations in production.

Keywords: Explainable AI; Stock Market Prediction; Risk Assessment; SHAP; Financial Machine Learning.

I. INTRODUCTION

Predicting asset prices is one of the harder problems in applied forecasting. The signal is weak, the noise is large, and any pattern that works tends to fade once people trade on it. Even so, machine learning has produced measurable out-of-sample gains. Gu, Kelly, and Xiu [1] found that neural networks and tree-based methods predicted cross-sectional stock returns better than linear benchmarks, and Fischer and Krauss [2] showed that LSTM networks could extract useful information from daily data on S&P 500 constituents. The difficulty is that these models are hard to read. Imagine a portfolio manager whose model has just turned bearish on bank stocks. If the model cannot say what drove that view, the manager cannot judge how far to trust it, and a risk officer has nothing to defend in front of a model validator. Supervisors make the same point in their own terms. The US guidance on model risk management, SR 11-7 [3], stresses conceptual soundness and effective challenge, while the EU AI Act [4] and the GDPR [5] place transparency and human oversight duties on some automated decisions.

Explainable AI (XAI) tries to close this gap. Two recent systematic reviews report that stock price prediction and credit management are among the most common finance tasks for which XAI is used [6], [7]. What remains unclear is how far explanations can be trusted on financial data, which is non stationary, heavily correlated, and easy to overfit. This review examines that question and offers a framework for using explanations responsibly.

A. Contributions

- 1) A theoretical grounding in predictability and risk measurement.
- 2) The mathematics of the main XAI methods and their failure modes in finance.
- 3) A comparison of representative applications to trend prediction and risk assessment.
- 4) A five-layer framework, together with regulatory and evaluation guidance.

II. LITERATURE SURVEY

A. Machine Learning in Market Forecasting

Before machine learning, return forecasting relied mostly on linear regressions and factor models. A typical factor model writes the return of asset i as

$$r_{i,t+1} = \alpha_i + \sum_k \beta_{i,k} f_{k,t+1} + \varepsilon_{i,t+1},$$

where the f_k are systematic factors such as market, size, and value [8]. Volatility was handled by ARCH and GARCH models [9], [10]. Machine learning relaxed the linearity assumption. Gradient-boosted trees [11], recurrent networks such as the LSTM [12], and more recently Transformers [13] can pick up nonlinear effects and interactions, and in several studies they improve forecasts [1], [2]. The improvements are usually modest in absolute terms, which is what one expects when most of the variation in returns is unpredictable.

B. Interpretability and Explainability

Authors use “interpretability” and “explainability” in slightly different ways, and Lipton [14] and Doshi-Velez and Kim [15] discuss the distinction at length. Here we call a model intrinsically interpretable if a person can follow how it works, as with a sparse linear model or a short rule list. We call an explanation post-hoc if it is produced afterwards for an otherwise opaque model. General surveys are available [16], [17], [18].

There is a real disagreement about which route is better. Rudin [19] argues that for high-stakes decisions one should use an interpretable model from the start instead of explaining a black box afterwards. The argument carries weight in finance, where simple models are often competitive and where a post-hoc explanation can only approximate what a complex model is doing.

C. Explanation Methods

Four families of methods appear repeatedly in the financial literature.

Feature attribution assigns each input a share of the prediction. SHAP [20] uses Shapley values, and TreeSHAP [21] makes the calculation fast for tree ensembles. LIME [22] fits a small interpretable model around one prediction. Integrated Gradients [23] and Grad-CAM [24] suit differentiable models, including those that read price charts as images. Global tools include permutation importance [25], partial dependence [26], and individual conditional expectation curves [27].

Attention-based explanation reads attention weights as a sign of which time steps or assets mattered. The Temporal Fusion Transformer [28] builds variable selection into a forecasting architecture for this reason. Jain and Wallace [29], however, found that attention weights are not reliably faithful explanations. Surrogate and rule-based explanation approximates a complex model with a simpler one, often a shallow tree, for audit purposes. Counterfactual explanation [30] asks what small change to the input would change the outcome. It suits risk decisions, such as what would bring a borrower under a risk threshold.

D. XAI for Stock Trend Prediction

The most common pattern is a gradient-boosted or random-forest model trained on technical, fundamental, and macro features and then explained with SHAP. Gu et al. [1] report that a small group of signal families, mainly momentum, liquidity, and volatility, accounts for much of the predictive power. That is an example of attribution yielding economic insight and not just a debugging aid. Recent work extends this to deep models. Muhammad et al. [31] built a deep network that classifies five types of trend on four stock indices and compared it with random forest, support vector machine, and logistic regression. They then used SHAP and LIME to pick the ten most influential features. Re-training on those ten lowered accuracy slightly while precision and recall rose, which is a useful reminder that explanation-driven feature pruning changes different metrics in different directions. Celik et al. [32] likewise layered explanation methods on top of machine-learning forecasts of financial time series.

For recurrent and convolutional networks, authors typically use gradient-based attribution or perturbation tests to find which lags or indicators matter, or they draw prices as images and apply Grad-CAM. These explanations depend on preprocessing and window length, so a single run should not be taken at face value. Transformer forecasters offer attention maps across time and assets, which help visualise dependencies but should be backed by a second method.

E. XAI for Risk Assessment

Credit risk is the most mature application. Bracke et al. [33] applied Shapley-based explanation to machine-learning default models in a Bank of England study. Bussmann et al. [34] combined Shapley values with correlation-network analysis for credit risk. Gramegna and Giudici [35] compared SHAP and LIME on credit data and found that the two can paint different pictures of the same model, which is a warning against relying on one tool. For market risk, explanations show which factors push predicted volatility or tail risk upward, such as realised volatility, credit spreads, or macro stress indicators, and help reconcile machine-learning forecasts with established models. At portfolio level, attribution extends the classical decomposition of risk to nonlinear models. In validation work, explanations support conceptual-soundness review, comparison with challenger models, and drift monitoring.

F. Earlier Surveys and the Gap

General XAI surveys [16], [17], [18] cover methods across all domains and say little about market data. Two systematic reviews focus on finance: one classifies 138 studies by financial task and method [6], and the other proposes a unified taxonomy and catalogues the methods in use [7]. These are valuable inventories. What we found missing, and what this review tries to supply, is a link between the explanation methods and the economic theory of what financial models can actually learn, together with an examination of how the methods fail on financial data and what to do about it.

III. PROBLEM DEFINITION

Three problems motivate this review.

- 1) Opacity. The models that forecast best are often the ones people can least inspect, yet financial users must be able to question a forecast before acting on it.
- 2) Reliability. Most explanation methods were designed for data that are independent and stable, such as images or tabular records. Financial data are neither. Features are highly correlated, relationships drift over time, and signal is weak relative to noise. An explanation that is faithful in one setting may mislead in another.
- 3) Fragmentation. The relevant material is spread across finance, machine learning, and regulation, with little guidance on how the pieces fit together in a working forecasting system.
- 4) Formal statement. Let x_t be the features available at time t (or a window of the last L observations), and let y_{t+h} be the target at horizon h , such as a return, a direction label, or a risk quantity. A model f_θ is fitted by minimising a regularised loss,

$$\hat{\theta} = \underset{\theta}{\operatorname{argmin}} \frac{1}{T} \sum_{t=1}^T \ell(f_\theta(x_t), y_{t+h}) + \Omega(\theta).$$

An explainer maps the model and an input to an explanation $e = E(f_\theta, x)$, most often a vector of attributions $\phi(x)$. The problem is to produce explanations that are faithful to the model, stable across inputs and time, and economically meaningful, in a setting where labels overlap when $h > 1$ and the data distribution shifts.

The review addresses four research questions.

- RQ1. What do theory and evidence say about what financial models can predict, and what does that imply for explanations?
- RQ2. How do the main XAI methods work, and where do their assumptions fail on financial data?
- RQ3. How have they been applied to trend prediction and risk assessment?
- RQ4. How should explainable forecasting systems be built, validated, and governed?

IV. METHODOLOGY / APPROACH

A. Review Approach

This is a narrative review. We selected sources for relevance and not through a systematic database search. The selection covered four kinds of material: the papers that define each explanation method; empirical studies that apply explanation to stock-return, trend, credit, or risk models; two recent finance-specific systematic reviews; and regulatory texts. We did not rerun any experiments, so every finding reported here comes from the published papers. Because the sample is not exhaustive, our conclusions describe patterns in the sources we read and should not be taken as a census of the field.

B. Predictability and Market Efficiency

The efficient market hypothesis [36] says that prices already reflect available information. If that were fully true, no model could consistently beat the market after costs. The evidence is mixed. Intermediate-horizon momentum [37] and factor premia [8] suggest that returns are partly predictable, whether as compensation for risk or because of behavioural and institutional frictions. Lo's adaptive markets hypothesis [38] reconciles the two views by treating efficiency as something that varies over time. Welch and Goyal [39] showed that many variables proposed for forecasting the equity premium fail out of sample.

This matters for explanation in two ways. If predictability is weak and unstable, then how stable an explanation is over time tells us something about whether the model found a lasting effect. And an attribution only has value if it can be tied to a plausible economic story.

Overfitting deserves special mention. Because researchers test many models and features, the best in-sample result is biased upward. Harvey, Liu, and Zhu [40] argue for higher statistical thresholds for new factors, and Bailey et al. [41], [42] formalise backtest overfitting and propose the deflated Sharpe ratio. An explanation of an overfitted model is simply an accurate description of noise.

C. Risk Measures

Let L be the portfolio loss over some horizon. Value-at-Risk at level α is the α -quantile of the loss distribution,

$$\text{VaR}_\alpha(L) = \inf\{\ell : \Pr(L \leq \ell) \geq \alpha\},$$

and Expected Shortfall averages the losses beyond that point,

$$\text{ES}_\alpha(L) = \frac{1}{1-\alpha} \int_\alpha^1 \text{VaR}_u(L) du.$$

Artzner et al. [43] defined coherent risk measures through four axioms: monotonicity, translation invariance, positive homogeneity, and subadditivity. ES satisfies them, while VaR does not always satisfy subadditivity, so it can fail to reward diversification. The Basel Committee's revised market-risk framework moved internal-model capital requirements to ES [44], which raises the importance of understanding what drives tail losses.

Volatility is usually modelled with GARCH(1,1) [10],

$$\sigma_t^2 = \omega + \alpha \varepsilon_{t-1}^2 + \beta \sigma_{t-1}^2.$$

This model is interpretable by design, since α is the reaction to new shocks and β the persistence of past variance. It therefore serves both as a benchmark for machine-learning volatility models and as a template for the kind of statement an explanation should be able to make. At portfolio level, Markowitz's variance formula [45] gives $\sigma^2 = w^\top \Sigma w$, and the Euler decom-

position splits it into position contributions $\text{RC}_i = w_i \partial \sigma / \partial w_i$

that add up to σ . This idea of contributions that add to the total

is what attribution methods carry over to nonlinear models. Table 1 summarises the main quantities.

Table 1. Risk quantities and what an explanation should convey.

Quantity	Interpretable benchmark	What to explain
Return forecast	Linear factor model	Which characteristics and states drive the forecast
Volatility	GARCH(1,1)	Contribution of recent shocks, persistence, macro stress
VaR	Historical or parametric VaR	Drivers of the quantile and its sensitivity
Expected Shortfall	Historical ES	Drivers of tail severity, not just frequency
Default probability	Logistic scorecard	Reasons per borrower; counterfactual recourse

D. Shapley-Value Attribution

Many popular explainers share one form, the additive feature attribution model [20],

$$g(z') = \phi_0 + \sum_{j=1}^M \phi_j z'_j, \quad z' \in \{0,1\}^M.$$

SHAP chooses the ϕ_j as Shapley values [46]. For a payoff function $v(S)$ defined on subsets S of the M features,

$$\phi_i = \sum_{S \subseteq N \setminus \{i\}} w_S [v(S \cup \{i\}) - v(S)],$$

$$w_S = \frac{|S|!(M - |S| - 1)!}{M!}.$$

Shapley values are the only attribution that satisfies four ax-ioms. Efficiency means the attributions add up to $f(x) E[f(X)]$. Symmetry gives equal credit to features that contribute equally. The dummy axiom gives zero to a feature that never changes the payoff. Additivity makes attributions

linear across games. Efficiency is what makes SHAP attractive for risk reporting, because the contributions sum exactly to the gap between one forecast and the average, in the same spirit as the Euler decomposition. The cost is computation: an exact answer needs 2^M coalitions, so practical tools rely on sampling [47], kernel-weighted regression (KernelSHAP), or tree-specific shortcuts [21].

The choice of payoff function matters a great deal. The inter-ventional version averages over the removed features independently of the kept ones, while the conditional version respects the correlations in the data:

$$v_i(S) = E_{X_S^-} [f(x_S, X_S^-)],$$

$$v_i(S) = E[f(X) | X_S = x_S].$$

The first can evaluate the model at unrealistic inputs. The second can give credit to features the model never uses simply because they correlate with ones it does [48], [49], [50]. Covert et al. [51] show that many explainers are variations on the same idea of removing features and summarising the effect.

In finance the issue is acute, because moving-average ratios, momentum, and oscillators are all built from the same price history. Reporting grouped attributions, for example one contribution for a “momentum family”, is often more meaningful than splitting credit among near-duplicates.

E. LIME, Gradients, and Global Tools

LIME [22] explains one prediction by fitting a sparse surrogate in a neighbourhood of the input,

$$\xi(x) = \underset{g \in G}{\operatorname{argmin}} L(f, g, \pi_x) + \Omega(g),$$

where π_x is a proximity kernel. The result depends on how the neighbourhood is built. For time series, perturbing each lag independently destroys autocorrelation and produces implausible samples, so perturbing contiguous segments is more defensible. Kernel width and sampling noise also make LIME vary between runs, which partly explains why SHAP and LIME can disagree [35].

For differentiable models, Integrated Gradients [23] integrates the gradient along a path from a baseline x' to the input,

$$IG_j(x) = (x_j - x'_j) \int_0^1 \frac{\partial f(x'_j + \alpha(x - x'))}{\partial x_j} d\alpha,$$

$$\sum_j IG_j(x) = f(x) - f(x').$$

The second line is the completeness property, which parallels Shapley efficiency. The baseline is a modelling decision: a window of zero returns, a historical-average window, and a calm-regime window each define a different reference and so give different explanations. For a sequence model the input is a lag-by-feature matrix, so the output is a heat-map that can be checked for plausibility. Grad-CAM [24] weights the final convolutional feature maps by pooled gradients and is used when price charts feed the model.

Partial dependence [26] averages predictions while varying a subset S of features,

$$f_S(x_S) = \frac{1}{n} \sum_{i=1}^n f(x_S, x_{-S}^i),$$

and ICE curves [27] show the unaveraged versions. Both assume the features in S are independent of the rest, which is rarely true of financial predictors. Permutation importance [25] measures how much the loss rises when one feature is shuffled. For time-ordered data, shuffling should be done in blocks so that serial structure unrelated to the feature is left intact.

F. Attention and Counterfactuals

Scaled dot-product attention [13] computes

$$\text{Attn}(Q, K, V) = \text{softmax} \left(\frac{QK^T}{\sqrt{d_k}} \right) V.$$

It is tempting to read each weight as the importance of a past day. But the output also depends on the value vectors, on multiple heads, and on later layers, and Jain and Wallace [29] found that very different weights can lead to the same prediction. We therefore treat attention as a hypothesis and confirm it by intervention, for instance by masking the heavily attended time step and measuring the change in output.

A counterfactual explanation looks for the smallest change in the input that produces a desired outcome [30],

$$x^{cf} = \underset{x}{\text{argmin}} d(x, x') \text{ s.t. } f(x') \geq \tau.$$

In finance the hard part is feasibility. Changing “price momentum” while holding “volatility” fixed may be impossible, since both come from the same price path. Causal reasoning [52],

[53] gives a language for saying which changes are admissible. Counterfactuals fit best where a decision threshold exists, such as credit and risk limits.

G. Attribution Drift

Financial relationships change over time, so the explanation itself changes. To make that measurable we use a simple attribution drift score. Let $\phi^-(k)$ be the mean absolute attribution vector over period W_k , and let ρ_S be the Spearman rank correlation. Then

$$D(W_1, W_2) = 1 - \frac{\rho_S(\phi^-(1), \phi^-(2))}{|\phi^-(1)|},$$

The score is near zero when the model leans on the same features in both periods and grows as the ranking of drivers changes. We offer it as a practical diagnostic and not as an established standard. Training and test samples should also be separated by purging and an embargo period when labels overlap [54].

H. Proposed Framework for Explainable Forecasting

Drawing on the theory above, we propose a five-layer framework (Table 2). It is a design and review checklist and has not been validated empirically.

Table 2. Layers of the proposed framework.

Layer	Core practices	Typical artefacts
1. Data and features	Point-in-time data, leakage controls, survivorship handling	Data lineage, feature dictionary
2. Modelling	Interpretable baseline plus complex challengers; time-respecting validation	Model registry, baseline comparison
3. Explanation	Several explainers; grouped attributions; stated baselines and value functions	Attribution reports, case files
4. Validation	Stability, fidelity, economic plausibility, multiple-testing corrections	Validation report, stress scenarios
5. Monitoring	Attribution-drift alerts, re-validation, versioned documentation	Drift dashboards, model cards

In practice the workflow runs as follows. Define the target and horizon and build features from information available at decision time. Split the data with purging and an embargo. Fit an interpretable baseline together with one or more complex models. Compute explanations out of sample with at least two methods, then test their stability across folds, seeds, and small perturbations. Check economic plausibility against the baseline's explanation, evaluate decision value net of costs [42], and monitor attribution drift once the model is live.

Six principles follow. Start with an interpretable baseline. Explain the quantity that matters, for example the difference between a machine-learning volatility forecast and a GARCH forecast, because raw attributions will mostly point to lagged volatility, which every risk manager already knows. Use more than one method. Group correlated features and state the value function used. Report how much attributions vary across re-trainings, and where possible attach predictive uncertainty such as Monte-Carlo dropout [55]. Finally, treat every explanation as a hypothesis about the model, to be tested against economic reasoning and out-of-sample evidence.

I. Evaluation Criteria

Evaluation of explanations is still the weakest part of the field. Several properties can be made precise. Fidelity asks whether features judged important really matter: remove the top-k features and see how far the prediction moves. Stability can be expressed as a local Lipschitz estimate [56],

$$S(x) = \max_{x' \in \mathcal{B}_\epsilon(x)} \frac{\|\phi(x) - \phi(x')\|_2}{\|x - x'\|_2}.$$

where small values mean nearby inputs receive similar explanations. Sparsity reflects how many drivers a person can realistically consider. Temporal consistency is captured by the drift score D. Hooker et al. [57] warn that removing features can confuse importance with distribution shift, which applies with extra force to financial series.

Explanations are only worth studying if the forecast itself is valid, so predictive evaluation comes first. Out-of-sample R2 against a benchmark forecast is a common summary,

$$R_{oos}^2 = 1 - \frac{\sum (y_t - \hat{y}_t)^2}{\sum (y_t - y_t^{bench})^2}$$

and differences between models can be tested with the Diebold–Mariano test [58]. For VaR, the Kupiec test [59] checks whether exceedances occur at the stated frequency and the Christoffersen test [60] checks whether they cluster. These tests judge the output, while explanations judge the mechanism, so the two are complementary. Table 3 organises the levels.

Table 3. Levels of evaluation.

Level	Example metrics	Question answered
Forecast	RMSE, directional accuracy, R2, Diebold–Mariano oos	Is the model predictive out of sample?
Risk	Kupiec and Christoffersen tests, ES back-tests	Are risk forecasts calibrated and independent?
Explanation	Fidelity, Lipschitz stability, sparsity, drift	Is the explanation faithful, stable, and concise?
Decision	Net Sharpe, deflated Sharpe, drawdown	Does acting on the model add value after costs?
Human	Task accuracy, time to detect errors	Do explanations improve human judgement?

V. RESULTS & DISCUSSION

A. Main Observations

Reading across the studies, we draw five observations.

- 1) Tree ensembles with SHAP dominate practice. The pairing is popular because it is fast, exact for trees, and works on tabular predictors. It is also where the correlated-feature problem bites hardest.
- 2) Explanations are most valuable as a validation tool.
- 3) The clearest benefits reported are catching leakage, pruning features, and spotting regime shifts, not delivering a story to end users.
- 4) Credit risk is ahead of market risk. Explanation is already tied to adverse-action reasons and regulatory review in credit, while market-risk work is thinner.
- 5) Methods disagree more than is often admitted. SHAP and LIME can rank the same features differently [35], and attention weights may not match gradient-based importance [29].
- 6) Evidence on decision value is scarce. In the studies we read, few show that explanation-guided decisions improve risk-adjusted returns or reduce losses.

B. Comparison with Previous Research

Table 4 sets representative studies side by side. The explanation approach and the stated takeaway differ widely, but a pattern appears: most studies explain which inputs mattered and few test whether the explanation was stable or useful.

Table 4. Representative studies and what they show.

Study	Task and model	Explanation approach	Takeaway
Gu et al. [1]	Return prediction with trees and neural networks	Variable-importance analysis	Momentum, liquidity, and volatility signals carry most weight
Fischer and Krauss [2]	Daily direction for S&P 500 stocks with LSTM	Descriptive look at selected stocks	LSTM extracted signal; explanation was informal
Muhammad et al. [31]	Five-class trend prediction, four indices, deep network	SHAP and LIME	Top-10 features cut inputs but moved metrics in different directions
Celik et al. [32]	Financial time-series forecasts, machine learning	XAI layered on forecasts	Explanation can be added to existing forecasters
Bracke et al. [33]	Default risk, machine learning	Shapley-based attribution	Explanations were adapted to supervisory use
Bussmann et al. [34]	Credit risk scoring	Shapley values with correlation networks	Network view added to feature attribution
Gramegna and Giudici [35]	Credit risk	SHAP versus LIME	The two methods can disagree

Compared with earlier surveys, our findings differ in emphasis. The finance-specific reviews [6], [7] organise the literature by task and method. We reach a complementary conclusion from the theory side: because predictability is weak and unstable, an explanation should be judged by its stability and economic meaning and not only by its visual clarity. This agrees with the caution raised in general XAI work about faithfulness [29], [57] and extends it to the specific hazards of financial data.

C. Comparison of Methods

Table 5 compares the methods covered here. The right-hand column is the one practitioners most often overlook, because each method fails in its own way on financial data.

D. Regulatory and Governance Implications

SR 11-7 [3] organises model risk controls around development and use, independent validation, and governance. A central idea is effective challenge, meaning critical review by informed and objective people. A validator who can inspect global attributions, local explanations for outliers, and their

Table 5. Explanation methods and their main caveats in finance.

Method	Scope	Principal caveat in financial use
TreeSHAP	Local, global	Credit splitting among correlated indicators; value-function choice
KernelSHAP	Local, global	Cost; unrealistic input combinations
LIME	Local	Neighbourhood and kernel sensitivity; perturbing time series
Integrated Gradients	Local	Baseline dependence
Grad-CAM	Local	Only for image-like encodings; coarse resolution
PDP / ICE	Global	Independence assumption; extrapolation
Permutation importance	Global	Correlated features; breaks serial structure
Attention weights	Local	Not guaranteed to be faithful
Counterfactuals	Local	Feasibility and causal validity

stability has far more to work with than one who sees only aggregate accuracy. The GDPR [5] limits decisions based solely on automated processing, but whether it creates a general right to explanation is debated [61]. The EU AI Act [4] takes a risk-based approach and lists systems that evaluate the creditworthiness of natural persons among its high-risk uses. Whether a given stock-prediction model falls inside a regulated category depends on what it does and where it is used, so institutions should check this case by case with legal advice. Model cards [62] offer a way to document intended use, data, performance, and limits, and an explainable-forecasting version could add the explainers used, their settings, and stability results (Table 6).

Table 6. Regulatory instruments and their implications for XAI.

Instrument	Focus	Implication for XAI
SR 11-7	Soundness, validation, governance of models	Explanations as evidence for effective challenge
GDPR Art. 22	Solely automated decisions about individuals	Meaningful information about logic; human review
EU AI Act	Risk-based obligations; credit scoring is high-risk	Transparency, documentation, human oversight
Basel market-risk rules	ES-based capital	Understand and validate tail-risk drivers

E. Challenges and Limitations

The first challenge is non-stationarity. An explanation describes the model fitted to the past, and when relationships drift, old attributions can mislead about the future. The second is correlated features. Technical indicators are highly collinear, so Shapley-based and permutation methods can split or misassign credit, and the choice between marginal and conditional value functions can change the result.

Explanations can also be unstable or manipulated. They can vary with sampling or retraining [63], and post-hoc explainers can be fooled on purpose [64], which is a governance risk if explanations are used for compliance. Overfitting is a re-lated danger: a plausible-looking explanation can lend false confidence to a strategy that only works in a backtest. And attributions describe what the model relies on, not how markets work, so reading a SHAP value as a causal driver is an error that appears often.

Other problems are practical. Signal-to-noise is low, so an explanation of a weak forecast may mostly explain noise. Studies differ in datasets, horizons, and cost assumptions, and look-ahead and survivorship bias are recurring threats. Different audiences, from quants to auditors, need different levels of detail. Finally, the supposed trade-off between accuracy and interpretability is often overstated [19], so an interpretable baseline should always be reported.

VI. CONCLUSION

Explainable AI has become a normal part of machine-learning-based financial forecasting. SHAP on tree ensembles is the dominant practice, and gradient- and attention-based methods are common for deep models. In credit risk, explanation is already tied to governance and regulation. The evidence base is still uneven, though. In the studies we read, explanations were rarely tested for fidelity, stability, or decision value, and financial data bring hazards of their own: regime change, collinearity, and overfitting. Responsible practice combines an interpretable baseline, validation that respects the order of time, more than one explainer, and a clear statement that attributions describe models, not markets. The five-layer framework proposed here is a starting point for putting that practice into a repeatable process.

VII. FUTURE SCOPE

Several directions look worth pursuing.

- 1) Causal and counterfactual explanation. Combining causal discovery with domain structure would move the question from “what did the model use” toward “what would change if we intervened”, with proper handling of confounding.
- 2) Monitoring of explanations. Treating attribution drift as a tracked quantity, alongside accuracy, and stress-testing whether strategies still rely on their intended drivers in different regimes.
- 3) Stable, correlation-aware attribution. Grouped and hierarchical Shapley variants, with uncertainty estimates for the attributions themselves.
- 4) Interpretable-by-design models. Sparse, monotonic, and additive neural architectures, and factor-structured networks consistent with asset-pricing theory.
- 5) Explanation of language and foundation models. Faithful explanations for LLM-based research agents and sentiment models, including where retrieved information came from.
- 6) Benchmarks and shared protocols. Open datasets, standard walk-forward splits, and metrics for fidelity, stability, and decision value designed for finance.
- 7) Studies with practitioners. Controlled experiments to find out whether explanations really help people make better decisions and catch errors sooner.
- 8) Regulatory alignment. Mapping XAI outputs to the documentation expected under model-risk, AI-governance, and consumer-protection regimes.

A. Limitations and benefits

This review has three limitations. It is narrative and not systematic, it relies on published reports without rerunning experiments, and the proposed framework has not yet been tested in practice. Its benefits over earlier work are that it joins the economic theory of predictability to the mathematics of explanation methods, covers trend prediction and risk assessment together, and turns the discussion into a concrete checklist that practitioners and validators can apply.

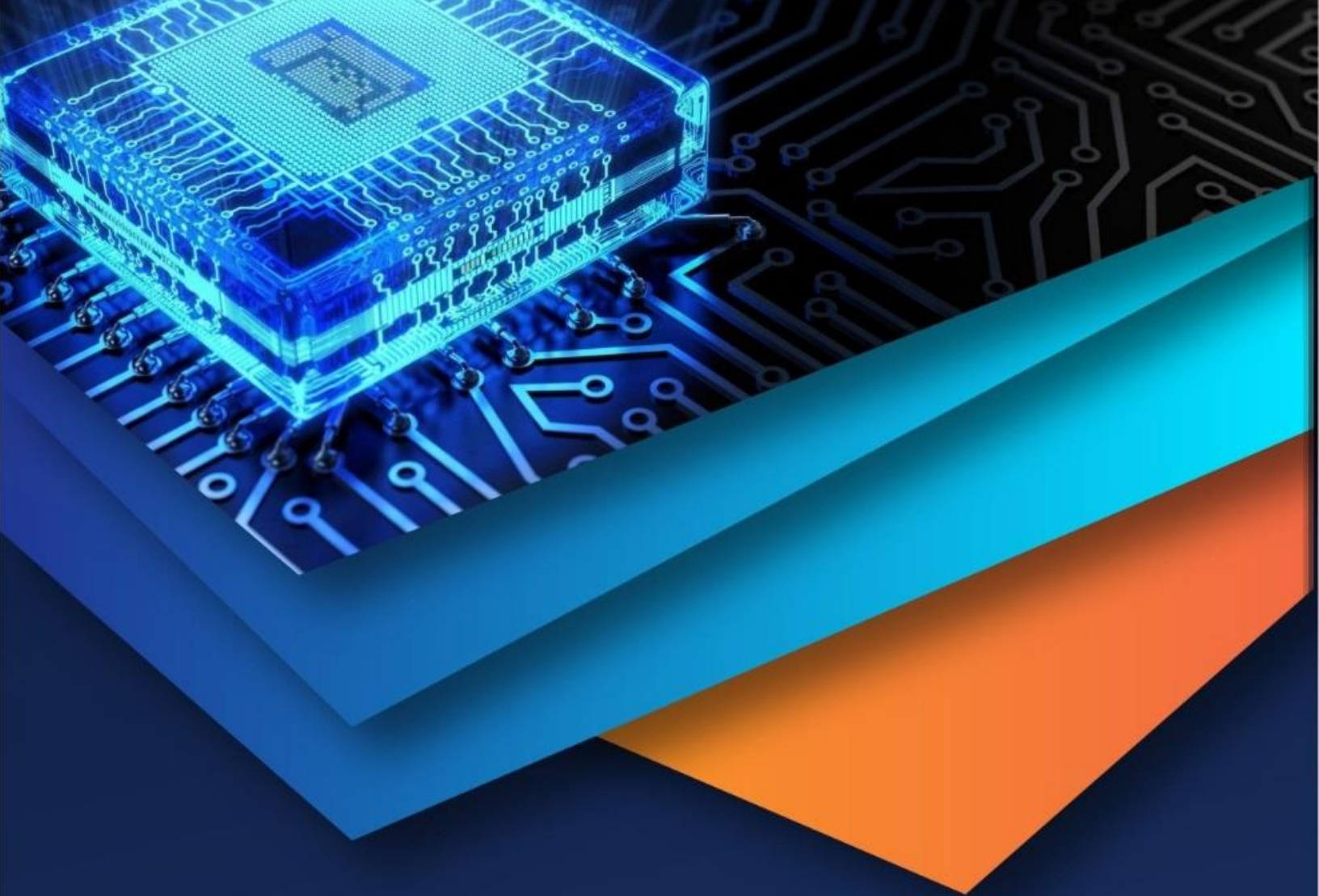
REFERENCES

- [1] Gu S, Kelly B, Xiu D. Empirical asset pricing via machine learning. *Rev Financ Stud.* 2020;33(5):2223–73.
- [2] Fischer T, Krauss C. Deep learning with long short-term memory networks for financial market predictions. *Eur J Oper Res.* 2018;270(2):654–69.
- [3] Board of Governors of the Federal Reserve System. Supervisory guidance on model risk management (SR 11-7). Washington (DC): Federal Reserve; 2011.
- [4] European Parliament and Council of the European Union. Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). *Off J Eur Union.* 2024.
- [5] European Parliament and Council of the European Union. Regulation (EU) 2016/679 (General Data Protection Regulation). *Off J Eur Union.* 2016;L119:1–88.
- [6] Černevičienė J, Kabašinskas A. Explainable artificial intelligence (XAI) in finance: a systematic literature review. *Artif Intell Rev.* 2024;57:216. doi:10.1007/s10462-024-10854-8.

- [7] Martins T, de Almeida A, Cardoso E, Nunes L. Explainable artificial intelligence (XAI): a systematic literature review on taxonomies and applications in finance. *IEEE Access*. 2024;12:618–29. doi:10.1109/ACCESS.2023.3347028.
- [8] Fama EF, French KR. Common risk factors in the returns on stocks and bonds. *J Financ Econ*. 1993;33(1):3–56.
- [9] Engle RF. Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation. *Econometrica*. 1982;50(4):987–1007.
- [10] Bollerslev T. Generalized autoregressive conditional heteroskedasticity. *J Econom*. 1986;31(3):307–27.
- [11] Chen T, Guestrin C. XGBoost: a scalable tree boosting system. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*; 2016. p. 785–94.
- [12] Hochreiter S, Schmidhuber J. Long short-term memory. *Neural Comput*. 1997;9(8):1735–80.
- [13] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. In: *Advances in Neural Information Processing Systems 30*; 2017. p. 5998–6008.
- [14] Lipton ZC. The mythos of model interpretability. *Commun ACM*. 2018;61(10):36–43.
- [15] Doshi-Velez F, Kim B. Towards a rigorous science of interpretable machine learning. arXiv:1702.08608. 2017.
- [16] Adadi A, Berrada M. Peeking inside the black-box: a survey on explainable artificial intelligence (XAI). *IEEE Access*. 2018;6:52138–60.
- [17] Guidotti R, Monreale A, Ruggieri S, Turini F, Giannotti F, Pedreschi D. A survey of methods for explaining black box models. *ACM Comput Surv*. 2018;51(5):Article 93.
- [18] Arrieta AB, Díaz-Rodríguez N, Del Ser J, Bénéto A, Tabik S, Bar-bado A, et al. Explainable artificial intelligence (XAI): concepts, taxonomies, opportunities and challenges toward responsible AI. *Inf Fusion*. 2020;58:82–115.
- [19] Rudin C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nat Mach Intell*. 2019;1(5):206–15.
- [20] Lundberg SM, Lee SI. A unified approach to interpreting model predictions. In: *Advances in Neural Information Processing Systems 30*; 2017. p. 4765–74.
- [21] Lundberg SM, Erion G, Chen H, DeGrave A, Prutkin JM, Nair B, et al. From local explanations to global understanding with explainable AI for trees. *Nat Mach Intell*. 2020;2(1):56–67.
- [22] Ribeiro MT, Singh S, Guestrin C. “Why should I trust you?”: explaining the predictions of any classifier. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*; 2016. p. 1135–44.
- [23] Sundararajan M, Taly A, Yan Q. Axiomatic attribution for deep networks. In: *Proceedings of the 34th International Conference on Machine Learning*; 2017. p. 3319–28.
- [24] Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-CAM: visual explanations from deep networks via gradient-based localization. In: *Proceedings of the IEEE International Conference on Computer Vision*; 2017. p. 618–26. Breiman L. Random forests. *Mach Learn*. 2001;45(1):5–32.
- [25] Friedman JH. Greedy function approximation: a gradient boosting machine. *Ann Stat*. 2001;29(5):1189–232.
- [26] Goldstein A, Kapelner A, Bleich J, Pitkin E. Peeking inside the black box: visualizing statistical learning with plots of individual conditional expectation. *J Comput Graph Stat*. 2015;24(1):44–65.
- [27] Lim B, Ank SÖ, Loeff N, Pfister T. Temporal fusion transformers for interpretable multi-horizon time series forecasting. *Int J Forecast*. 2021;37(4):1748–64.
- [28] Jain S, Wallace BC. Attention is not explanation. In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*; 2019. p. 3543–56.
- [29] Wachter S, Mittelstadt B, Russell C. Counterfactual explanations without opening the black box: automated decisions and the GDPR. *Harv J Law Technol*. 2017;31(2):841–87.
- [30] Muhammad D, Ahmed I, Naveed K, Bendechache M. An explainable deep learning approach for stock market trend prediction. *Heliyon*. 2024;10(21):e40095. doi:10.1016/j.heliyon.2024.e40095.
- [31] Celik TB, Ican Ö, Bulut E. Extending machine learning prediction capabilities by explainable AI in financial time series prediction. *Appl Soft Comput*. 2023;132:109876. doi:10.1016/j.asoc.2022.109876.
- [32] Bracke P, Datta A, Jung C, Sen S. Machine learning explainability in finance: an application to default risk analysis. Bank of England Staff Working Paper No. 816. London: Bank of England; 2019.
- [33] Bussmann N, Giudici P, Marinelli D, Papenbrock J. Explainable machine learning in credit risk management. *Comput Econ*. 2021;57(1):203–16.
- [34] Gramegna A, Giudici P. SHAP and LIME: an evaluation of discriminative power in credit risk. *Front Artif Intell*. 2021;4:752558.
- [35] Fama EF. Efficient capital markets: a review of theory and empirical work. *J Finance*. 1970;25(2):383–417.
- [36] Jegadeesh N, Titman S. Returns to buying winners and selling losers: implications for stock market efficiency. *J Finance*. 1993;48(1):65–91.
- [37] Lo AW. The adaptive markets hypothesis: market efficiency from an evolutionary perspective. *J Portf Manag*. 2004;30(5):15–29.
- [38] Welch I, Goyal A. A comprehensive look at the empirical performance of equity premium prediction. *Rev Financ Stud*. 2008;21(4):1455–508.
- [39] Harvey CR, Liu Y, Zhu H. and the cross-section of expected returns. *Rev Financ Stud*. 2016;29(1):5–68.
- [40] Bailey DH, Borwein J, López de Prado M, Zhu QJ. Pseudo-mathematics and financial charlatanism: the effects of backtest overfitting on out-of-sample performance. *Not Am Math Soc*. 2014;61(5):458–71.
- [41] Bailey DH, López de Prado M. The deflated Sharpe ratio: correcting for selection bias, backtest overfitting, and non-normality. *J Portf Manag*. 2014;40(5):94–107.
- [42] Artzner P, Delbaen F, Eber JM, Heath D. Coherent measures of risk. *Math Finance*. 1999;9(3):203–28.
- [43] Basel Committee on Banking Supervision. Minimum capital requirements for market risk. Basel: Bank for International Settlements; 2019.
- [44] Markowitz H. Portfolio selection. *J Finance*. 1952;7(1):77–91.
- [45] Shapley LS. A value for n-person games. In: Kuhn HW, Tucker AW, editors. *Contributions to the theory of games, vol II*. Princeton (NJ): Princeton University Press; 1953. p. 307–17.
- [46] Štrumbelj E, Kononenko I. Explaining prediction models and individual predictions with feature contributions. *Knowl Inf Syst*. 2014;41(3):647–65.
- [47] Janzing D, Minorics L, Blöbaum P. Feature relevance quantification in explainable AI: a causal problem. In: *Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics*; 2020. p. 2907–16.
- [48] Sundararajan M, Najmi A. The many Shapley values for model explanation. In: *Proceedings of the 37th International Conference on Machine Learning*; 2020. p. 9269–78.
- [49] Kumar IE, Venkatasubramanian S, Scheidegger C, Friedler S. Problems with Shapley-value-based explanations as feature importance measures. In: *Proceedings of the 37th International Conference on Machine Learning*; 2020. p. 5491–500.
- [50] Covert I, Lundberg SM, Lee SI. Explaining by removing: a unified framework for model explanation. *J Mach Learn Res*. 2021;22(209):1–90.
- [51] Pearl J. Causality: models, reasoning, and inference. 2nd ed. Cambridge: Cambridge University Press; 2009.



- [52] Peters J, Janzing D, Schölkopf B. Elements of causal inference: foundations and learning algorithms. Cambridge (MA): MIT Press; 2017.
- [53] López de Prado M. Advances in financial machine learning. Hoboken (NJ): Wiley; 2018.
- [54] Gal Y, Ghahramani Z. Dropout as a Bayesian approximation: representing model uncertainty in deep learning. In: Proceedings of the 33rd International Conference on Machine Learning; 2016. p. 1050–9.
- [55] Alvarez-Melis D, Jaakkola TS. On the robustness of interpretability methods. arXiv:1806.08049. 2018.
- [56] Hooker S, Erhan D, Kindermans PJ, Kim B. A benchmark for interpretability methods in deep neural networks. In: Advances in Neural Information Processing Systems 32; 2019.
- [57] Diebold FX, Mariano RS. Comparing predictive accuracy. J Bus Econ Stat. 1995;13(3):253–63.
- [58] Kupiec PH. Techniques for verifying the accuracy of risk measurement models. J Deriv. 1995;3(2):73–84.
- [59] Christoffersen PF. Evaluating interval forecasts. Int Econ Rev. 1998;39(4):841–62.
- [60] Wachter S, Mittelstadt B, Floridi L. Why a right to explanation of automated decision-making does not exist in the General Data Protection Regulation. Int Data Priv Law. 2017;7(2):76–99.
- [61] Mitchell M, Wu S, Zaldivar A, Barnes P, Vasserman L, Hutchinson B, et al. Model cards for model reporting. In: Proceedings of the Conference on Fairness, Accountability, and Transparency; 2019. p. 220–9.
- [62] Ghorbani A, Abid A, Zou J. Interpretation of neural networks is fragile. Proc AAAI Conf Artif Intell. 2019;33(1):3681–8.
- [63] Slack D, Hilgard S, Jia E, Singh S, Lakkaraju H. Fooling LIME and SHAP: adversarial attacks on post hoc explanation methods. In: Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society; 2020. p. 180–6.



10.22214/IJRASET



45.98



IMPACT FACTOR:
7.129



IMPACT FACTOR:
7.429



INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Call : 08813907089  (24*7 Support on Whatsapp)