Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Design and Methodology of an Explainable Hybrid Multi-modal AI Framework for NIFTY 50 Stock Market Prediction

Amit Kumar Verma

amitverma.researchwork@gmail.com

Suresh Gyan Vihar University

Himanshu Vasnani

Suresh Gyan Vihar University

Deep Shikha

Arrivo EDU Private Limited

Research Article

Keywords: stock market prediction, multimodal artificial intelligence, hybrid learning, technical indicators, fundamental analysis, financial sentiment, macroeconomic indicators, FinBERT, late fusion, stacking, SHAP, explainable AI

Posted Date: November 25th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11210836/v1

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

Additional Declarations: No competing interests reported.

Abstract

The prediction of stock markets is challenging due to the presence of heterogeneous, nonlinear, time-varying and partially observable information.Current literature is scattered between modalities, classes of models and evaluation protocols. It does not offer a common approach to model selection, information fusion, leakage control and interpretability.In this paper, we propose a fully specified and explainable hybrid multimodal AI architecture for NIFTY 50 index in India, combining four different sources of information, namely technical information from the market (41 indicators), fundamentals of the firms, sentiment scores of the financial news obtained through FinBERT, and market stress measures derived from india vix. Instead of assuming an architecture beforehand, the framework specifies a common pool of nineteen AI algorithmic modelsfor systematic evaluation of each modalityunder a 5-fold walk-forward expanding-window protocol, and the framework defines model selection using an eight-metric (RMSE, MAE, MAPE, R^2, Adj-R^2, Directional Accuracy, Theil's U_2, Spearman IC) weighted Composite Score together withDiebold-Mariano significance testing; selected models are then fused using a novel three-stage fusion engine, which sequences graph-based, attention-based and SHAP-aligned weighting, and SHAP is used to determine which information sources are driving predictions, and whether these vary across market regimes.This proposed architecture does not only capture the magnitude of the prediction made by the model, but it also answers the fundamental question of why the model made that prediction and which information source was more useful in this prediction.

1. Introduction

What is the source of information that is used in financial markets? History of various professions, stock exchange data, technical indicators, data on the financial situation of companies, macroeconomic data, public opinion and news. It's all important and the literature of the efficient markets has always been demanding when it comes to handling it; prices should react quickly to information; and systematic prediction should be difficult [1]. That hasn't been sustained in a perfect manner. The non-linear relationships may be found in large predictor datasets that aren't modeled by simpler methods, and in some cases, those relationships will result in out-of-sample improvements [2]. That has resulted in forecasting being detached from simply price-related models and now, deep learning has grown to be the main body of it – LSTM and GRU models were employed to capture the temporal dependencies [3] and hybrid convolutional-recurrent-attention models with different types of signals [4]. One restriction still exists. Technical analysis is completely devoid of any information concerning the company fundamentals, news text or macroeconomic background.It's not immediately apparent in the price in the financial text. That is why the domain-specific language models are important and FinBERT was born: financial language is not general domain language, and FinBERT can capture the sentiment signal which can not be captured by numerical features [5]. Now it is scalable, FNSPID matches with price in huge datasets, but no combination of all four modalities, with fundamentals and macro data exists so far [6]. However, the mixture of these resources is extremely challenging. Technical Data is being shown on a daily basis. Fundamentals arrive quarterly. There are no regular releases for the macro. If it happens, news breaks. If one simply tosses everything in with no care, the differences in the statistical properties

of their respective components just vanish — which is why there has been a renewed interest in multimodal work with explicit fusion mechanisms, temporal-alignment and modality-specific representation-learning [7]. The third one is related to the issue of interpretability. Finally, complex ensembles, and deep networks in general, are just hard to audit, and SHAP can help audit it by providing a common additive framework for attributing predictions to features [8]; but feature level explanations can't tell whether or not the prediction was based on technical, fundamental, sentiment or macroeconomic data, or how the weights of these vary in volatile markets.Here in one place the three problems are solved. The proposal is an explainable hybrid multimodal AI framework that adopts a pipeline of data acquisition and alignment, 4 modality-specific modelling, explainability and evaluation, towards the objective 1 of the overall research programme. This design has the following consequences: it is a four-modality design rather than a design based on four add-on modalities; in the context of this design, a hybrid late-fusion strategy is taken, which allows modelling the modalities according to their statistical characteristics; modality-level SHAP analysis is possible; the evaluation protocol is tailored to minimize the look-ahead bias among the modalities, that is, between the single-modality and the early and hybrid designs.The conceptual architecture, the 4 modality data design, the modality-specific modelling logic, the fusion strategy and the explainability layer based on SHAP on which the rest of the objectives are realised is also presented. The parts of the model are empirically instantiated in two additional papers: one is a second paper that reports the best-performing model on each of the four data sources (technical, fundamental, sentiment, macroeconomic (India VIX)), in the common pool of candidate models and 5-fold walk-forward selection procedure described in Section 5; and one is a third paper that fuses the four best-performing models, one from each of the four data sources, through the hybrid structure described in Section 6, reporting Directional Accuracy as primary empirical result and the remaining Composite Score components (Section 8.3), the narrower papers that follow can thus be read as tests of the hypotheses outlined here, rather than as independent investigations.

2. Literature Review and Research Gap

2.1 From Efficient Markets to Machine Learning

It is usually assumed that the markets are efficient, meaning prices adjust quickly to new information, and that the possibility of predictable price behavior should be hard to uncover [1] and this is indeed what subsequent market-efficient tests have found [15]. This is not what machine learning is challenging, but rather, whether there is incremental prediction information in relationships that are nonlinear and high dimensional that is not provided by simpler models. Trees and neural networks are able to capture that kind of interaction in asset pricing. But regularisation and out-of-sample testing should still be used, since as flexibility increases, there is a true overfitting risk. [2].

2.2 Technical and Temporal Modeling

Technical indicators: Moving averages, RSI, MACD, Bollinger Bands, volatility, volume are all methods of organizing the raw price history. It is a sequence that an LSTM network would be able to handle. Memory

is retained over long time periods and the vanishing-gradient problem of previous recurrent models is eliminated [9] and this was later generalized [18]. The results of the deep-learning analysis show how pervasive LSTM and GRU are in financial forecasting [3]. But, no matter how developed a technical model is, it can only be safe as long as there is no change of regime.

2.3 Fundamental Information

The fundamental variables such as profitability, leverage, liquidity, valuation, growth are updating only periodically and the availability of the value on a point in time is crucial: only after the public release date of the value, not at the end of the fiscal period, shall the value enter the model. Tree-based models are appropriate for such a heterogeneous tabular predictor, and nonlinear models are able to pick up information from a large predictor set.[2].

2.4 Financial Sentiment and Domain Language Models

Prices may not fully reflect information in financial text. FinBERT is one solution: it is a transformation of transformer-based language representations to finance-specific jargon [5]. Researchers have shared-task benchmarks through which they can assess such sentiment extraction [14]. On the data side, FNSPID provides millions of time-aligned price-news records, enabling large scale integration [6] and recent multimodal studies already show benefits from combining technical and textual sentiment data with attention-based late fusion.[7].

2.5 Macroeconomic Information

The rates, the inflation, the industrial activity, the exchange rate, the commodities, the credit conditions, and the volatility – all the macroeconomic variables that matter, and equity returns are not the same during tightening, inflation shocks, and stress. However, there is an issue with that. These observations will be published at a later date and may be amended after publication. The reason the model is not able to look ahead here is because of a release-date alignment and lag structures.

2.6 Multimodal Fusion

So to combine sources – which do not seem to be alike at all – and to not lose what makes each source useful – it's the whole point of multimodal learning. This survey is comprehensive, and categorises the field into the three design families: early feature level fusion, intermediate representation level fusion and late decision level fusion. As noted in [7] it still faces problems of temporal misalignment, modality imbalance, missing data and generalization across markets. In the case of fusion, it's not something to be left to chance, and there's evidence to support that: a modality-level attention study using technical, investor-sentiment and earnings-call data found that all three modalities actually contributed to prediction in a late fusion architecture. [10].

2.7 Explainable AI in Financial Prediction

There's no room for lack of explainability here. Researchers, investors, risk managers and regulators always end up looking at predictions at some point, and SHAP provides them with a point of access: it

provides additive feature attributions that have consistency properties making comparisons meaningful [8]. At two junctures, this framework is based on SHAP: within each modality-specific model, to understand which variables contribute to the final call; and in the fusion model, to understand the contribution of each modality.

2.8 Research Gap

A companion systematic review of over 50 studies on AI-based stock-market-prediction studies published in highly-reputable international journals 2018–2025 [20] reveals five parallel gaps in this literature: most studies combine at most two types of data together, there is limited principled consideration of model choice given the data type, there is limited testing of model-joining with respect to single models, SHAP is only applied in single data-single model settings, and newer tools (large language models, trading agents, graph-based models, privacy-preserving learning) are only evaluated individually. The literature summarized in this section indicates five related gaps in the current framework, in addition to the one outlined in the accompanying review. Second, much of the research is limited to combining two or three modalities, typically price and technical data with sentiment, and excluding the incorporation of detailed fundamentals or macroeconomic variables in a unified design. Second, temporal-alignment procedures vary a lot from study to study, and are often left undocumented. Third, no one has standardized when the early fusion should be compared to the specialized late fusion. A fourth gap is that explainability is typically added to a single model as opposed to the entire multimodal decision making process. Fifth and last, with different prediction targets, prediction horizons, markets, data periods, and prediction metrics, it is difficult to compare performance across papers. In order to fill these gaps, the proposed framework attempts to do so in a concerted manner, as all the four sources of information are treated as first class modalities, data timing, specialized models, fusion, explainability and evaluation are all described within the same research design.

3. Research Objectives, Questions and Conceptual Framework

3.1 Objective

The objective of this paper is to design and formalize a unified, explainable hybrid multimodal AI architecture for Nifty 50 stock market prediction. The framework brings together the four kinds of information: technical, fundamental, sentiment, and market stress represented by india vix. Each one keeps its own statistical character rather than getting smoothed into a single generic input.

3.2 Research Questions

RQ1: How can heterogeneous technical, fundamental, sentiment, and market-stress information be structured within a unified multimodal architecture for NIFTY-50 prediction?

RQ2: How can modality-specific model selection and leakage-controlled temporal validation be incorporated into a common multimodal prediction framework?

RQ3: How can out-of-fold predictions from modality-specific models be integrated through a hybrid late-fusion mechanism while preserving the distinct contribution of each modality?

RQ4: How can feature-level, modality-level, and regime-wise explainability be incorporated into the architecture to provide insight into why a prediction is generated and which information sources contribute to that decision?

3.3 Framework Architecture

The proposed framework is organized as an end-to-end architecture for NIFTY-50 prediction.It begins with four heterogeneous information modalities : technical market information, firm-level fundamentals, financial-news sentiment, and market stress represented by the India VIX. Each modality is processed separately to preserve its underlying characteristics before being integrated at the fusion stage.The architecture first performs data preprocessing, temporal alignment, and feature representation for each modality. A common pool of nineteen predictive models is then defined to support modality-specific benchmarking and model selection within the proposed architecture. The architecture defines a leakage-controlled walk-forward validation procedure for modality-specific model selection, followed by the generation of out-of-fold predictions for subsequent fusion.The resulting modality-specific outputs are then passed to the proposed hybrid late-fusion stage. The proposed fusion design combines stacking, attention-based weighting, and SHAP-aligned weighting to construct an integrated representation for the final prediction task. An extended graph-based fusion pathway is alsoas an alternative architectural variant for subsequent empirical evaluation.Explainability is incorporated throughout the architecture rather than being treated only as a post-hoc analysis. Feature-level, modality-level, and regime-wise SHAP analyses are defined to examine the contribution of individual features and information sources to the model’s prediction. The complete architectural flow is summarized in Table 1 and Fig. 1.

Table 1
Framework components, modality-specific representations, model-selection strategy, and fusion pathway.

Component	Technical	Fundamental	Sentiment	Market Stress
Input	Technical market information	Firm-level fundamental information	Financial-news information	India VIX and related market-stress information
Representation	Sequential feature representation	Tabular feature representation	FinBERT-based sentiment representation	Time-aligned sequential representation
Model selection	Selection from Common 19-model pool	Selection from Common 19-model pool	Selection from Common 19-model pool	Selection from Common 19-model pool
Validation	Walk-forward expanding window	Walk-forward expanding window	Walk-forward expanding window	Walk-forward expanding window
Output to fusion	Out-of-fold prediction / representation	Out-of-fold prediction / representation	Out-of-fold prediction / representation	Out-of-fold prediction / representation
Fusion role	Modality-specific contribution	Modality-specific contribution	Modality-specific contribution	Modality-specific contribution

4. Data Construction and Temporal Alignment

4.1 Study Universe and Target

The proposed framework is designed for NIFTY-50 prediction and can be extended to other liquid equity indices. It supports next-day return regression and direction classification. The primary target and study period will be specified in the subsequent empirical study. All predictor information must be available at or before the decision cutoff on day t.

4.2 Technical Data

Technical modality consists of a whole set of technical indicators based on market information through back-looking time windows. These indicators reflect various aspects of price action, momentum, trend, volatility, volume and other market dynamics. The architecture maintains time correlation in the technical information and uses only information up to the specified decision deadline. A detailed description of the indicator specifications and their empirical configuration is given in the following implementation and validation study.

4.3 Fundamental Data

The fundamental modality represents firm-level financial and valuation information obtained from publicly available disclosures. Observations are aligned using their public release dates rather than

fiscal-period dates to maintain point-in-time consistency.

4.4 Sentiment Data

Sentiment modality refers to the information contained in financial news, financial announcements and related articles. The language representation component is FinBERT to obtain sentiment information and dense representations. News is aggregated based on the specified decision cutoff, hence no future information is allowed to be part of the feature set

4.5 Market-Stress Data and Synchronization

The market-stress modality is represented by the India VIX, which captures market-based expected volatility. India VIX observations are aligned with the NIFTY-50 trading calendar while respecting their temporal availability.

4.6 Data Leakage Controls

The framework incorporates leakage controls throughout the data pipeline. Scaling and feature selection are restricted to the training window. Fundamental data use public release dates, news is restricted to pre-cutoff information, and technical variables use only backward-looking observations. Training, validation, and test periods remain chronologically separated.

4.7 Proposed Feature Groups

Table 2
Proposed feature groups by modality, temporal frequency, and primary data-leakage risk.

Modality	Frequency	Primary leakage risk
Technical	Daily	Temporal leakage from future observations
Fundamental	Quarterly/annual	Point-in-time leakage
Sentiment	Event/daily	Timestamp leakage
Market Stress	Daily	Release and availability timing

Within the proposed architecture, the four modalities are considered as four different information streams. The streams are subjected to modality-specific pre-processing, temporal alignment and representation prior to the modeling stage. The detailed set of features, study period, data sources, model configuration and quantitative results shall be reserved for the following empirical implementation and validation study.

5. Modality-Specific Modeling

The proposed framework does not assume that a single predictive model is equally suitable for all information sources. Instead, each modality is processed through a common pool of predictive models,

allowing the model-selection stage to account for differences in data structure, dimensionality, temporal behavior, and information content.

Figure 2 presents the proposed leakage-controlled model-selection workflow. Each modality is processed independently, evaluated using the same temporal validation structure, and passed through the model-selection mechanism.

5.1 Common AI Model Pool

Instead of specifying the type of architecture for the AI model for each modality, 19 AI models are generically specified on each of the six paradigms of different architectures and the best model for each modality is selected by the model selection mechanism. The nineteen AI models are divided into six major modeling paradigms: linear models (Linear Regression, Lasso, ElasticNet, Ridge, Ridge [Tuned]); ensemble models (Extra Trees, Random Forest, XGBoost, Ensemble [Top-K], LightGBM); recurrent models (Gated Recurrent Unit [GRU], Long Short-Term Memory [LSTM], Bidirectional LSTM [Bi-LSTM] and CNN-LSTM); attention-based models (Transformer [13]); and statistical models (Naïve Random Walk [RW] and Prophet). The evaluation is done on the same nineteen-model pool for each modality with a 5-fold walk-forward expanding window validation procedure. Each model is given the current training period and tested on the next validation period and the training period is extended till the next fold. This is to avoid look-ahead bias and for temporal ordering. The Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE), coefficient of determination (R^2), Adjusted R^2, Directional Accuracy, Theil's U_2 and Spearman Information Coefficient (IC) are used to evaluate the performance of the model. Each of these 8 metrics is then weighted and combined to make a weighted Composite Score which is averaged across each of the 5 folds. The highest Composite Score is specified as the criterion for selecting the best model for each modality; the differences in the scores of competing models are tested pairwise using the Diebold–Mariano test and controlling for multiple testing using false-discovery-rate (FDR) procedures. The modality-specific models selected are then used to make out-of-fold predictions which acts as leakage-free inputs to the hybrid late fusion mechanism.

5.2 Technical Branch

The technical branch receives the engineered technical feature vector (Section 4.2), represented as a rolling sequence for sequence-based candidates and a flat vector for tabular/linear candidates; all nineteen common pool AI algorithms are processed on it. The candidate models from the common pool are evaluated using the proposed walk-forward procedure. The architecture does not assume that recurrent or deep-learning models are necessarily preferable for technical information. Instead, the model-selection mechanism allows the characteristics of the technical representation to determine the most suitable predictive model.The resulting modality-specific model output is subsequently used to generate out-of-fold predictions for the fusion stage.

5.3 Fundamental Branch

The fundamental branch receives the structured financial and valuation representation described in Section 4.3 and processes it through the same common AI model pool and leakage-controlled validation protocol. The architecture allows different model families to be considered according to the characteristics of the fundamental information, without assuming a particular model in advance. The modality-specific model is subsequently selected through the proposed model-selection mechanism.

5.4 Sentiment Branch

The sentiment branch uses FinBERT to transform financial-news documents into contextual representations and sentiment-related outputs [5]. These representations are aggregated over the defined daily time window to construct a fixed-length sentiment feature representation. The resulting representation, rather than the raw text, is then processed through the common AI model pool under the proposed model-selection protocol. FinBERT serves as the language representation component, while the subsequent predictive model is determined through the modality-specific model-selection mechanism.

5.5 Market-stress Branch

The market-stress branch receives the India VIX-based representation described in Section 4.5, together with permissible lagged information where defined by the proposed temporal-alignment procedure. The resulting representation is processed through the common AI model pool under the same leakage-controlled walk-forward validation protocol. The architecture does not assume a particular model family for this modality; instead, the appropriate modality-specific model is determined through the proposed model-selection mechanism. The resulting out-of-fold predictions are then passed to the hybrid late-fusion stage.

5.6 Model Selection Rationale

The four modalities differ in their temporal structure, frequency, dimensionality, and representation. Therefore, the architecture allows each modality to follow its own model-selection pathway while maintaining the same overall evaluation protocol.The framework does not require all four modalities to use the same model family. Instead, a modality-specific model is selected through the defined model-selection mechanism. The resulting selected models then generate out-of-fold predictions, which serve as inputs to the hybrid late-fusion mechanism described in Section 6.

6. Hybrid Fusion Architecture

6.1 Late Fusion by Stacking

The primary fusion mechanism is late fusion through stacking: each base learner is the top-performing model selected for its modality in Section 5, trained on only that modality's feature vector, producing out-of-sample predictions z_t = [yhat^T_t, yhat^F_t, yhat^S_t, yhat^M_t] that a meta-learner g(.) maps to the

final prediction yhat_t = g(z_t); candidate meta-learners include regularized linear/logistic regression, gradient boosting, and a shallow MLP.

Meta-learner training must use predictions generated without fitting the corresponding base model on the same observations, using the same 5-fold walk-forward expanding-window out-of-fold predictions as Section 5.1, with the test period untouched until all choices are frozen.

6.2 Attention-Based Fusion

An optional extension replaces scalar-prediction stacking with modality embeddings h^m_t, projected into a common latent dimension and passed to an attention layer whose weights α^m_t (Σ_m α^m_t = 1) yield the fused representation h_t = Σ_m α^m_t h^m_t—a data-dependent weighting mechanism, though attention weights should not be read as causal importance; SHAP is retained for post-hoc attribution.

6.3 Proposed Hybrid Architecture

Table 3
Five-stage proposed hybrid architecture: operation and output at each stage, from independent modality preprocessing to SHAP explanation.

Stage	Operation	Output
1	Independent modality preprocessing	X^T, X^F, X^S, X^M
2	Specialized base models	yhat^T, yhat^F, yhat^S, yhat^M
3	Out-of-sample prediction generation	Z
4	Meta-learning / attention fusion	yhat
5	SHAP explanation	Feature and modality attribution

6.4 Extended Three-Stage Fusion Engine

It is important to note that the above stacking and attention mechanism are not treated as alternatives, but are both used in the proposed extended fusion engine as its final architecture. The modality embeddings first go through a graph-based fusion stage, where an inter-modality graph is used to allow a GNN to learn cross modal dependencies instead of treating modalities as independent; the result is then passed through attention based fusion (Section 6.2), which highlights the more informative modalities per observation; a third stage, explainability aligned weighted fusion, uses modality importance derived from SHAP to align the modality fusion weights with the explainability layer (Section 7) so that the two accounts are consistent. The output is a single fused latent vector, which is connected to a prediction head (MLP, LSTM, Transformer) that generates a return, direction, risk score, or trading signal, as measured by directional accuracy, the ROC-AUC, F1, RMSE, Sharpe ratio and information ratio; the feedback is propagated end-to-end through the graph and attention stages.

7. Explainability Layer

7.1 SHAP Foundations

SHAP is based on an additive explanation framework $f(x) = \varphi_0 + \Sigma_j \varphi_j$, where each feature j's value is a contribution relative to a reference value, and can be used for both global and local explanations. For candidates that are trees, there's a shortcut: the Exact SHAP is computed directly, rather than using the slower model-agnostic estimator.[19].

7.2 Modality-Specific SHAP

Which variables are predominant in each branch? Modality-specific SHAP does that: momentum, volatility, RSI, MACD, moving averages or volume (technical); profitability, leverage, liquidity, valuation or growth (fundamental); document-level variables (sentiment); or rates, inflation, commodities or volatility (macro).In the subsequent empirical implementation, absolute SHAP values can be aggregated over the test set to obtain a global importance measure for each modality.

7.3 Fusion-Level SHAP

Now, SHAP is also applied at the meta-learner, with its four modality-level predictions as the features to explain. That's a numeric value of the real contribution of each modality to the final product. Perhaps its major contribution to explainability is that it translates the question of “which variable matters?” to “which information source matters for this decision?”

7.4 Regime-Specific Explainability

How to partition the test time into regimes? Any rule that has an economic impact is good; trend/realized-volatility classification or volatility-state model, either will do — as long as they're fixed before the analysis begins. Following that split, In the subsequent empirical implementation, SHAP distributions can be analyzed across bull, bear, sideways, low-vol and high-vol time periods, with case studies of notable policy events, earnings releases, and geopolitical events providing further context.

7.5 Interpretation Caveat

SHAP tells you who your model is pointing to, but not which is the economically more important cause, which is important to stress. If a feature has a high SHAP value, there is only one conclusion to be drawn: the feature had an influence on the model output for the distribution of the features used for the reference. But it does not imply that the feature brought the market move and the manuscript should maintain that separation.

8. Proposed Experimental Design and Evaluation Protocol

8.1 Proposed Baselines

Baselines include a naive/random-direction benchmark, a historical benchmark, and the following statistical models: ARIMA, GARCH, along with the naive and Prophet candidates already in the common

pool (Section 5.1), single-modality baselines using the best pool candidate on each of the four feature vectors (Sections 5.2–5.5), an early-fusion model on concatenated multimodal features, a simple average of modality predictions, the proposed late-fusion stacking model (Section 6.1), and attention-based fusion (Section 6.2) and the extended graph/attention/SHAP-aligned fusion engine (Section 6.4).

8.2 Proposed Data Splitting Protocol

No random shuffling of the split, it is all chronological. The main protocol used is 5-fold walk-forward expanding-window cross validation (Section 5.1), where a window is used to train each model, the next blocks are set aside and predicted, and the window is expanded to include realized outcomes, with 5-fold cross validation to ensure that each candidate model sees the same out-of-sample blocks. A static split (e.g., 70/15/15 train/validation/test) can also be reported as a robustness check, and is not the underlying split used for the main results.

8.3 Proposed Evaluation Metrics

The regression metrics reported are MAE, RMSE, MSE, R-squared and directional accuracy; the classification metrics reported are accuracy, precision, recall, F1 and ROC-AUC; and an economic evaluation under a given trading rule (return, volatility, Sharpe ratio, maximum drawdown, turnover and transaction-cost adjusted performance) is also reported because financial prediction is a decision oriented problem.

For the technical/directional-prediction pool (Section 5.1), the regression formulation is compared on eight metrics—RMSE, MAE, MAPE, R^2, Adjusted R^2, Directional Accuracy, Theil's U_2, Spearman IC—unified into a weighted Composite Score so no model wins by optimizing one metric alone; the classification formulation is compared on accuracy plus the Section 8.4 battery

8.4 Statistical Testing

A point estimate is not final on a forecast comparison. Comparing accuracy across different loss functions using a Diebold-Mariano test [12] helps, but is not sufficient by itself — issues of multiple comparisons remain.

In practice, that means to run a certain set of tests (not just one). Pairwise Diebold-Mariano tests get combined with Newey-West correction, to handle serial dependence. Bonferroni and Holm-Bonferroni corrections keep the family wise error rate in check for all comparisons made. Cohen's h checks for practical significance beyond the p values, while cluster-aware permutation importance finishes it off by removing those features that don't stand up to resampling.

9. Proposed Outcome Structure

Apart from the four research questions, this framework is supposed to provide three outputs: a prediction for the next day trend, a score of confidence and a SHAP-based explainability layer for all four modalities. That combination produces an actionable outlook, too, as well as a risk-aware reliability

measurement and a degree of transparency into why a forecast turned out as it did, instead of an opaque number returned.

10. Discussion

The prediction of stock market is not a simple time series problem here, since the variables (technical, fundamental, textual, macroeconomic) are different in terms of their frequencies, dimensionalities, noise, and arrivals (which are modular in nature); a modular architecture is easier to defend than one generic encoder for all. There is also a diagnostic pay-off, as with each modality alone before fusion, it is possible to see whether the modality contributes little or nothing, whether it is useful in certain regimes, or whether it outcompetes the others. Deliberately, point-in-time alignment is not a by-product of the pre-processing, but rather a methodological variable, since it touches upon the central challenge of the multimodal finance sector: portability - a model may be good due to accidentally providing updated or post-cutoff information. Explainability thus comes in two forms: on the one hand, a modality could have a high SHAP value at the level of fusion, but at the level of individual features, none of the features has a high value (due to the compression of the encoder, such as FinBERT, which blurs the signal of each feature); on the other hand, a modality could have only a few features with high SHAP value, but the modality as a whole is technically dominant.

All of this doesn't change the actual constraints. It is robustly non-stationary and has structural breaks along all the architectures. Even with the most advanced models, public data sets are subject to survivorship and coverage biases and basic and macro data remain revision-prone in the absence of a vintage database for comparison. SHAP values are not estimates of causal effects—rather, they are attributions. There's a framework that gives researchers this degree of freedom to select and that's why it's important to have rules for evaluation preregistered. This is confirmed by a recent survey [7] which shows that there are still missing-modality situations, and that data streams in the real world are sometimes temporally misaligned, or even lost without known cause, or modality imbalance.

11. Conclusion

In this paper, we propose a design and methodology of an explainable hybrid multi-modal AI framework for NIFTY 50 stock market prediction that integrates four distinct information sources (technical market information, firm fundamental information, financial sentiment information, and macroeconomic information) by combining modality-specific models with late fusion and SHAP at the feature level, and at the modality level, to overcome the central technical challenges of multimodal financial forecasting: temporal misalignment, look-ahead bias, scaling leakage, heterogeneous frequencies, and overfitting.

The basic idea is straightforward: financial data does not have the same predictive power in different markets, and the predictive power of financial data varies in the time scale, from technical data reflecting short-term market dynamics, to fundamental data reflecting longer-term market dynamics, to sentiment data reflecting the sudden shifts in market dynamics, to macro data reflecting market regimes. What

preserves these differences, and does not flatten them, and still learns how to combine them at the end is a specialized hybrid architecture.

12. Scope of the Present Study

This study focuses on the design and formalization of the proposed explainable hybrid multi-modal AI framework for NIFTY 50 stock market prediction. Empirical implementation, model training, quantitative performance evaluation, and comparison with baseline models are outside the scope of the present study and are reserved for subsequent empirical validation. Proposed framework: overall information flow over the four data modalities, structured process for temporal alignment and leakage control. It further incorporates modality-specific feature representation and a common AI model pool for systematic model selection. The framework also includes a walk-forward validation mechanism to provide leakage-free inputs for the subsequent multi-modal fusion and an out-of-fold prediction mechanism. The stacking, attention-based weighting, and SHAP-aligned weighting are all used at the fusion stage to fuse the information from the individual modalities. Lastly, the framework adds feature level, modality level and regime-wise SHAP components to facilitate more transparent interpretation of the model prediction.

Declarations

Funding: This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors. The authors did not receive any funding for this work.

Clinical trial registration: Not applicable. This study is not a clinical trial.

Ethics, Consent to Participate, and Consent to Publish declarations: Not applicable.

Data Availability : All data generated or analyzed during this study are included in this published article and its supplementary information files.

References

1. Malkiel, B. G., & Fama, E. F. (1970). Efficient capital markets: A review of theory and empirical work. *The Journal of Finance*, *25*(2), 383–417. https://doi.org/10.1111/j.1540-6261.1970.tb00518.x
2. Gu, S., Kelly, B., & Xiu, D. (2020). Empirical asset pricing via machine learning. *The Review of Financial Studies*, *33*(5), 2223–2273. https://doi.org/10.1093/rfs/hhaa009
3. Sezer, O. B., Gudelek, M. U., & Ozbayoglu, A. M. (2020). Financial time series forecasting with deep learning: A systematic literature review: 2005–2019. *Applied Soft Computing*, *90*, 106181. https://doi.org/10.1016/j.asoc.2020.106181
4. Huang, W., & Zhang, Y. (2023). & others. Deep learning for financial time series prediction: A state-of-the-art review of standalone and hybrid models. Computer Modeling in Engineering & Sciences,

139(1), 187–224. https://doi.org/10.32604/cmes.2023.031388

5. Araci, D. (2019). FinBERT: Financial sentiment analysis with pre-trained language models. arXiv:1908.10063.
6. Dong, Z., Fan, X., & Peng, Z. (2024). FNSPID: A comprehensive financial news dataset in time series. Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, 4918–4927. https://doi.org/10.1145/3637528.3671629
7. Liu, Y., Lin, Y., Xu, C., Wang, Q. J., Zhao, J., Guan, Z., Zhao, W., & Xu, G. (2026). Multi-modal fusion approaches for stock market prediction: A comprehensive survey. *Pattern Recognition*. https://doi.org/10.1016/j.patcog.2026.114384
8. Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, 30.
9. Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, *9*(8), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735
10. Hajek, et al. (2025). Multimodal financial sentiment for stock return prediction. *Procedia Computer Science*, *270*, 582–591. https://doi.org/10.1016/j.procs.2025.09.177
11. Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 785–794. https://doi.org/10.1145/2939672.2939785
12. Diebold, F. X., & Mariano, R. S. (1995). Comparing predictive accuracy. *Journal of Business & Economic Statistics*, *13*(3), 253–263. https://doi.org/10.1080/07350015.1995.10524599
13. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, 30.
14. Cortis, K., Freitas, A., Daudert, T., Hürlimann, M., Zarrouk, M., Handschuh, S., & Davis, B. (2017). SemEval-2017 Task 5: Fine-grained sentiment analysis on financial microblogs and news. Proceedings of the 11th International Workshop on Semantic Evaluation, 519–535.
15. Fama, E. F. (1991). Efficient capital markets: II. *The Journal of Finance*, *46*(5), 1575–1617.
16. Engle, R. F. (1982). Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation. *Econometrica*, *50*(4), 987–1007.
17. Bollerslev, T. (1986). Generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics*, *31*(3), 307–327.
18. Hochreiter, S., Bengio, Y., Frasconi, P., & Schmidhuber, J. (2001). Gradient flow in recurrent nets: The difficulty of learning long-term dependencies. In A Field Guide to Dynamical Recurrent Neural Networks.
19. Lundberg, S. M., Erion, G., Chen, H., DeGrave, A., Prutkin, J. M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N., & Lee, S. I. (2020). From local explanations to global understanding with explainable AI for trees. *Nature Machine Intelligence*, *2*, 56–67.

20. Verma, A. K., & Vasnani, H. (2026). A systematic review of artificial intelligence-based stock market prediction and the convergence toward multimodal explainable frameworks. SGVU International Journal of Convergence of Technology and Management, 12(2), 207–216. E-ISSN: 2455–7528.

Figures

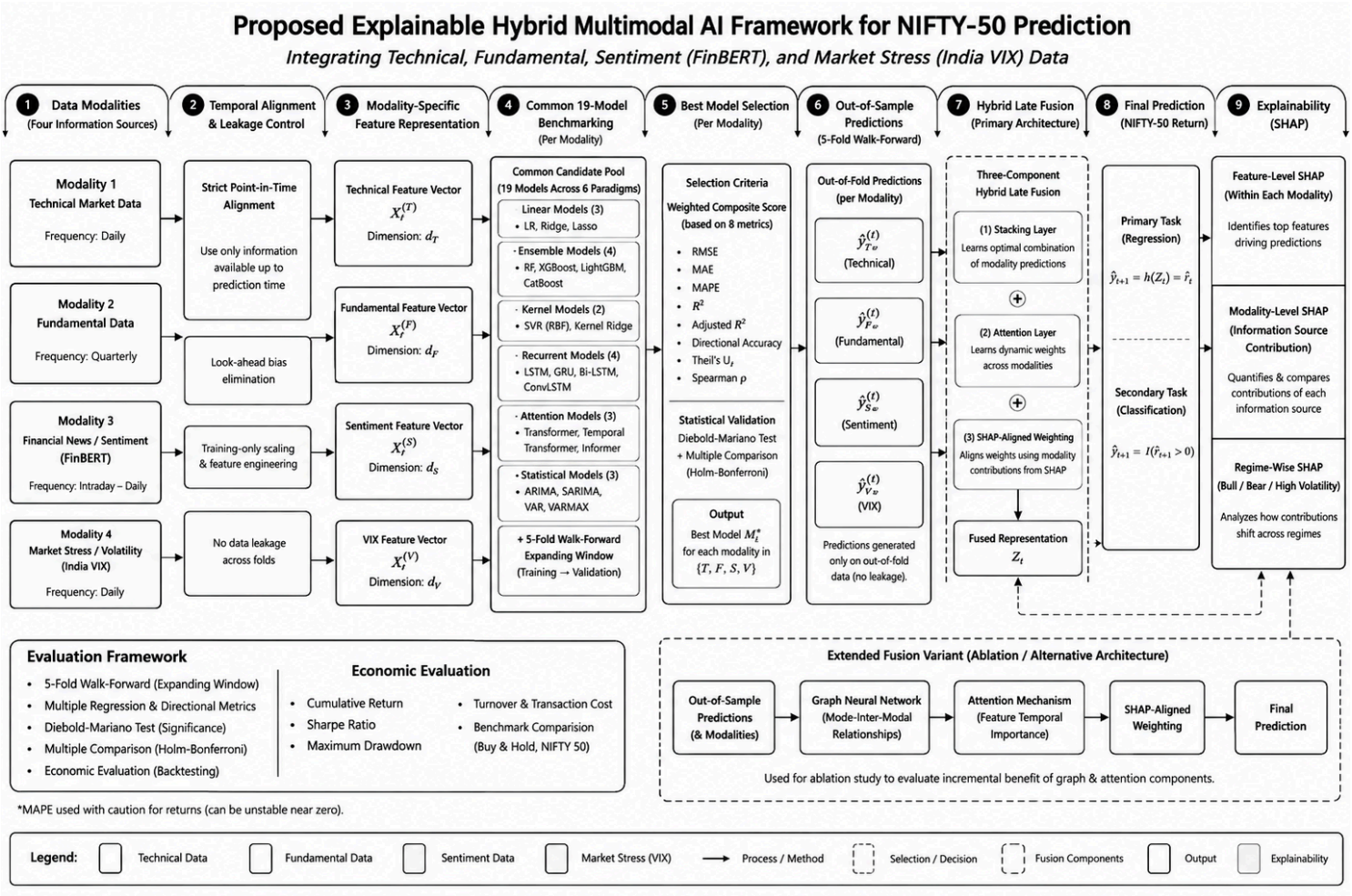

Figure 1

summarizes the end-to-end architecture of the proposed framework, from multimodal data acquisition, temporal alignment, and modality-specific representation to predictive-model selection, out-of-fold prediction, hybrid late fusion, NIFTY-50 forecasting, and multi-level SHAP-based explainability.

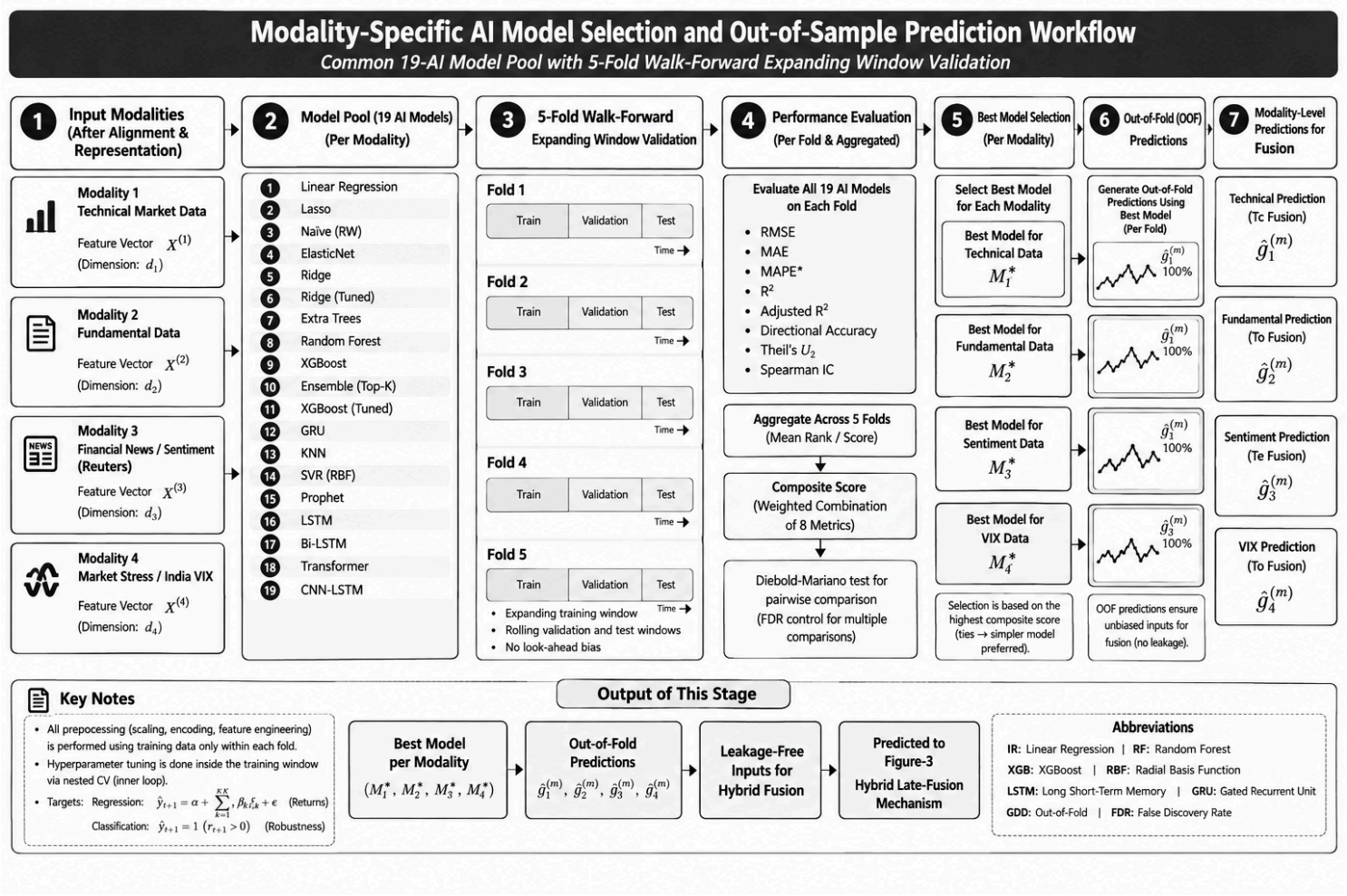

Figure 2

summarizes the leakage-controlled model selection process, from modality-specific AI benchmarking and walk-forward validation to best-model selection and generation of out-of-fold predictions for hybrid late fusion.

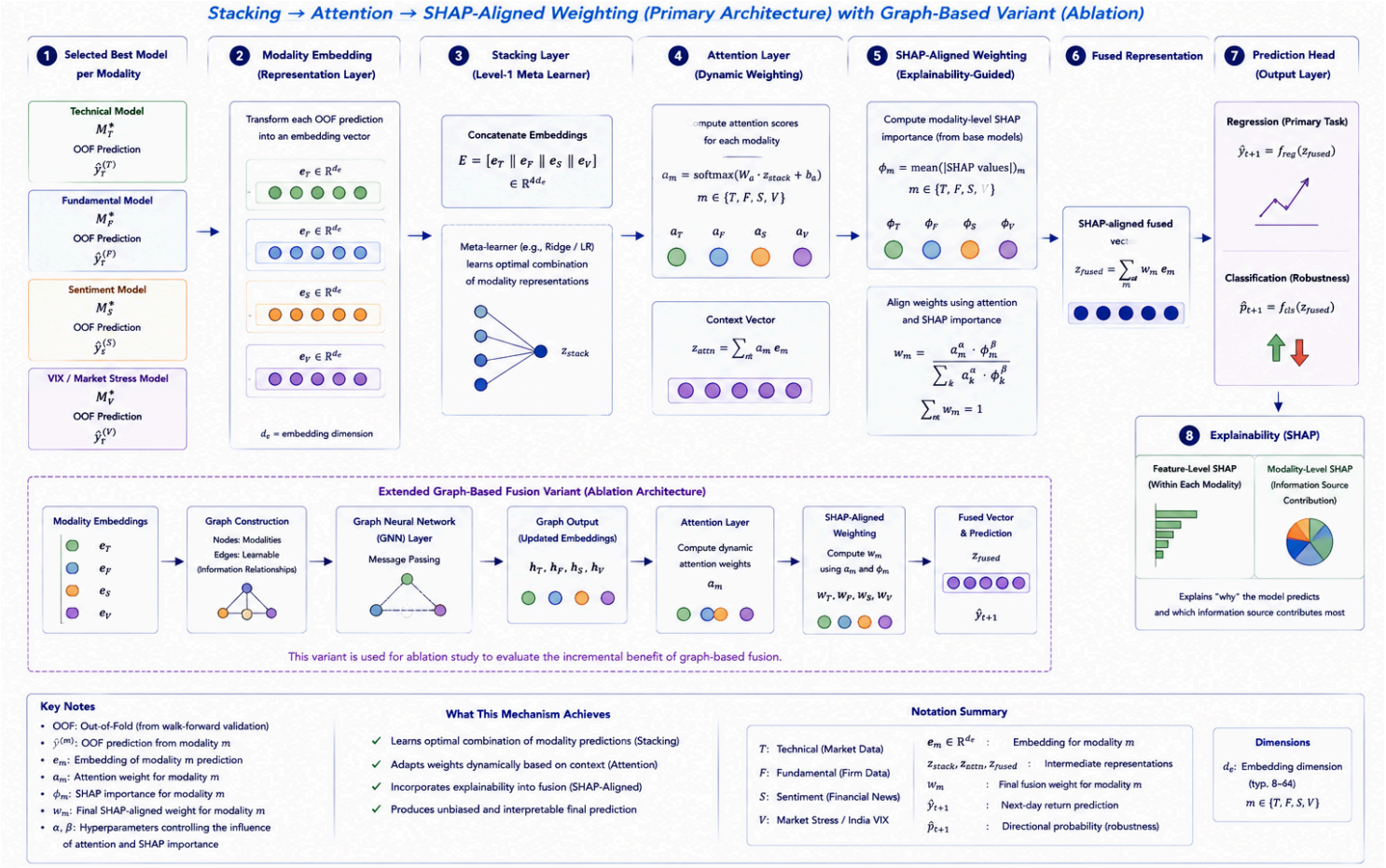

Figure 3

shows an overview of the proposed hybrid late fusion approach, where modality-specific out-of-fold predictions are combined using stacking, attention weighting and SHAP guided fusion, and graph-based fusion is proposed as an extended architectural variant for subsequent empirical evaluation.

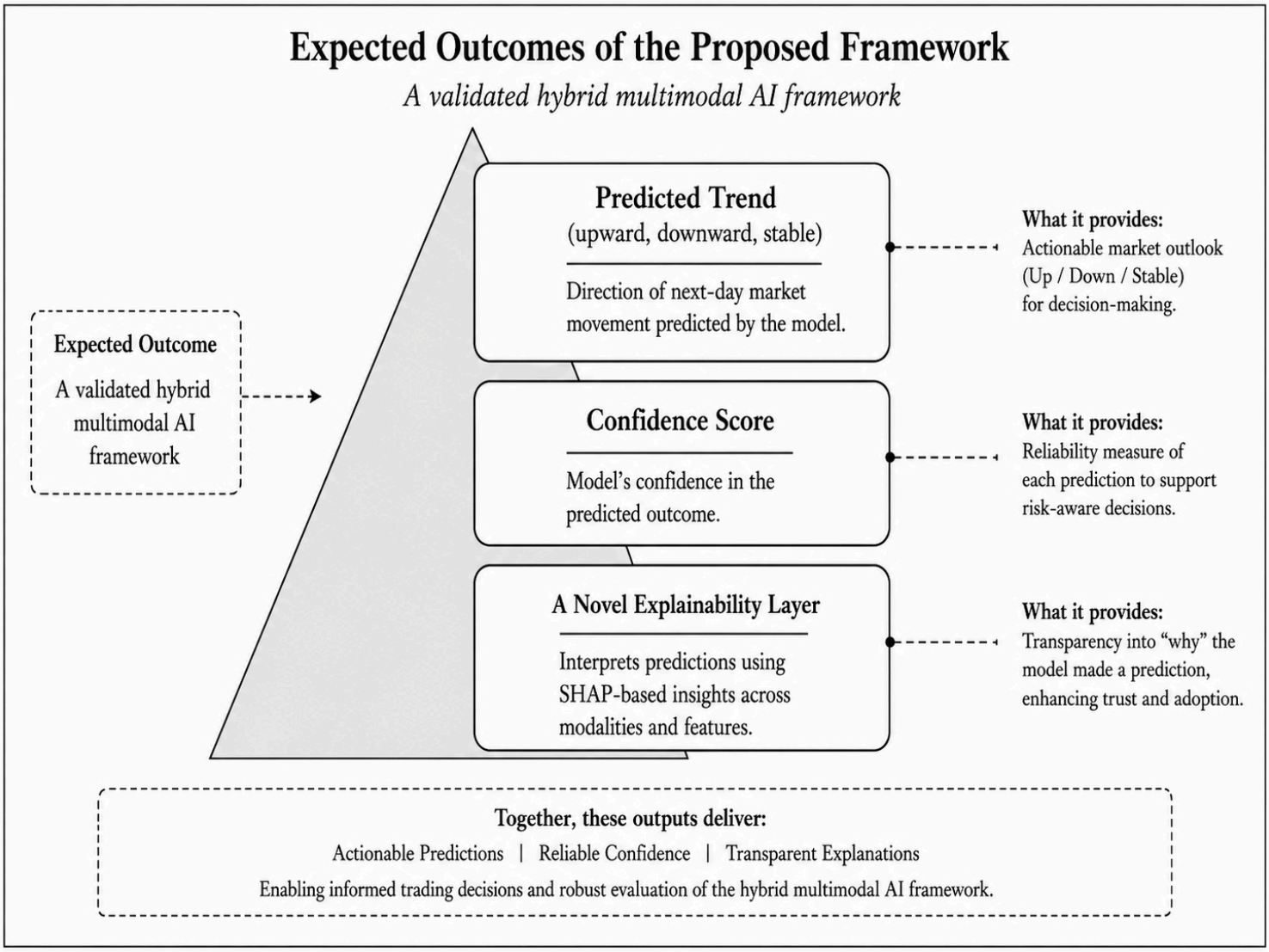

Figure 4

Predicted trend, confidence score, and SHAP-based explainability layer — together, what this framework is expected to deliver, giving actionable, reliable, transparent predictions.