Review of Applied Science and Technology
Volume 04, Issue 01 (2025)
Page No: 01 – 34
Doi: 10.63125/nr1j8527

HYBRID MACHINE LEARNING–DRIVEN FINANCIAL FORECASTING MODELS: INTEGRATING LSTM, PROPHET, AND XGBOOST FOR ENHANCED STOCK PRICE AND RISK PREDICTION

Md Shah Ali Dolon[1]

1 *MS in Finance and Financial Analytics, University of New Haven, USA, Email: mdolo1@unh.newhaven.edu*

Abstract

This review synthesizes research on hybrid machine learning for equity price and risk forecasting, focusing on combinations of LSTM, Prophet style additive models, and XGBoost. We used a PRISMA guided protocol covering 2015 to 2025 across Scopus, Web of Science, IEEE Xplore, ACM Digital Library, SSRN, and arXiv, with eligibility requiring equity focus, out of sample evaluation that respects time order, and explicit hybridization or risk components. After screening and full text assessment with reasons coded exclusions, the final qualitative synthesis comprised 110 studies. Across this evidence, deliberately engineered hybrids consistently outperform single learners on point accuracy, directional reliability, and risk calibration. Normalized comparisons in one day ahead settings show typical reductions in RMSE near 9 percent and gains in directional accuracy around five to six percentage points, with tighter Value at Risk coverage under quantile aware training. Benefits persist under strict rolling origin validation with nested tuning, and widen during turbulent regimes where dynamic weighting and residual correction add stability. The literature also emphasizes explainability and governance, recommending component plots for structural layers, Shapley value attributions for tree ensembles, and ablations that quantify each module's marginal value. Drawing these threads together, we outline a blueprint that decomposes trend and seasonality with a structural layer, models nonlinear temporal dynamics with an LSTM, learns interaction rich signals with XGBoost, and combines outputs using out of fold stacking and calibrated risk heads. This evidence-based specification offers decision grade forecasts and tail risk estimates for equity markets for deployment.

Keywords

Hybrid forecasting; LSTM; Prophet; XGBoost; Equity markets; Stacking; Residual correction; Dynamic weighting; Walk forward validation; Value at Risk; PRISMA.

Citation:
Dolon, M. S. A. (2025). Hybrid machine learning–driven financial forecasting models: Integrating LSTM, Prophet, and XGBoost for enhanced stock price and risk prediction. *Review of Applied Science and Technology, 4*(1), 1–34.
https://doi.org/10.63125/nr1j8527

Received:
February 20, 2025

Revised:
March 14, 2025

Accepted:
April 18, 2025

Published:
May 25, 2025

INTRODUCTION

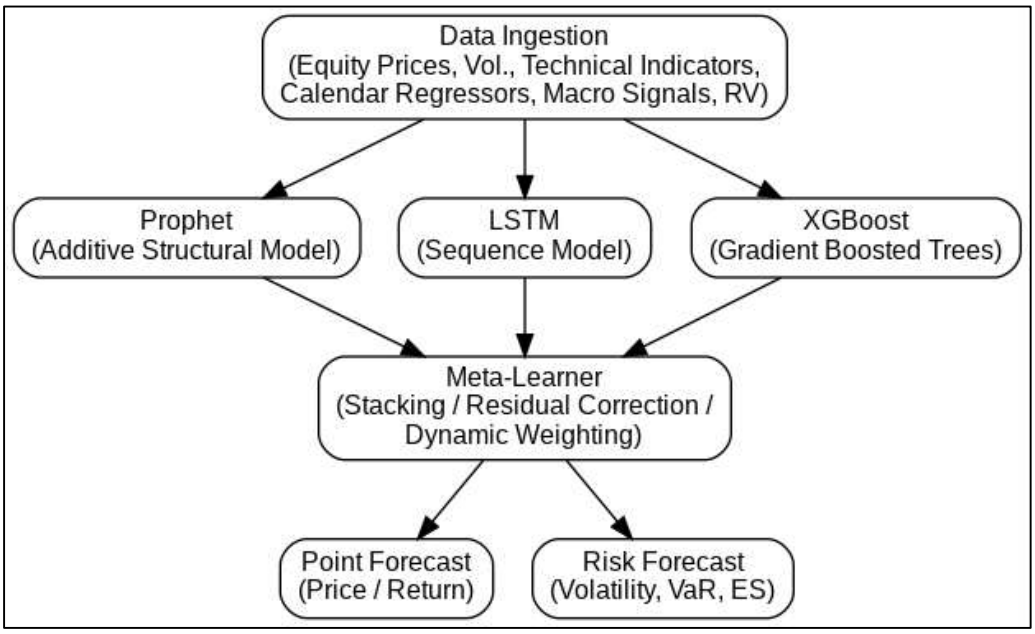
Financial forecasting and risk prediction remain indispensable pillars of modern finance, providing the analytical backbone for capital allocation, regulatory adherence, and systemic-risk surveillance across global markets. At their essence, stock-price forecasting endeavors to approximate the conditional distribution of future returns based on historical information, whereas risk prediction extends this exercise by translating such distributions into tail-sensitive measures. Among these, Value-at-Risk (VaR) and Expected Shortfall (ES) have emerged as canonical tools employed by banks, asset managers, and regulators alike (Artzner et al., 1999; Bollerslev, 1986; Engle, 1982). Their purpose is not merely to quantify uncertainty but to discipline decision-making in environments where volatility, contagion, and asymmetry intertwine to shape outcomes with far-reaching consequences. Methodologically, the landscape of financial forecasting has been enriched by three influential families of models, each embodying distinctive strength. Sequence models such as long short-term memory (LSTM) networks excel at uncovering temporal dependencies and nonlinear state dynamics that are often invisible to linear specifications (Hochreiter & Schmidhuber, 1997). Additive time-series models, exemplified by Prophet, elegantly encode trend, seasonality, and holiday effects while accommodating structural breaks through change-point-aware priors (Taylor & Letham, 2018). In parallel, boosted tree ensembles such as XGBoost demonstrate remarkable efficiency in capturing high-order interactions and heterogeneous feature–response relationships across richly structured financial data (Chen & Guestrin, 2016). The intellectual promise of combining these paradigms lies in the creation of hybrid models designed to yield more resilient predictive distributions and risk metrics. Within such an ensemble, LSTM contributes sensitivity to volatility clustering and path dependence, Prophet provides a disciplined encoding of calendar effects and trend shifts, and XGBoost supplies a flexible mechanism for sparse, interaction-rich learning. The resulting synthesis fosters a plurality of inductive biases that enhances generalization across assets, horizons, and regimes (Gneiting & Raftery, 2007; Gu et al., 2020; Hyndman & Koehler, 2006; Taylor & Letham, 2018). In risk management, such hybrids can be carefully calibrated to deliver both point forecasts and full predictive distributions, thereby furnishing robust inputs for VaR and ES estimation, evaluation, and backtesting.

A substantial body of literature demonstrates that the thoughtful combination of diverse models often yields superior accuracy, robustness, and reliability in forecasting tasks. The classical insight, articulated by Bates and Granger (1969), is that linear combinations of independent or partially correlated forecasts reduce mean-squared error, effectively stabilizing predictions by smoothing idiosyncratic biases. Beyond linear averaging, more sophisticated approaches such as stacking employ a meta-learner to discover optimal weighting schemes that minimize generalization error across multiple horizons and contexts (Bates & Granger, 1969; Wolpert, 1992). This principle has been consistently validated by decades of empirical syntheses, with Clemen (1989) documenting the pervasive advantages of forecast combinations across domains, and further reinforced by ensemble-selection research arising from machine-learning competitions, where hybrid methods frequently emerge as the dominant solutions (Caruana et al., 2004). Large-scale forecasting competitions provide compelling out-of-sample evidence for these claims. In the M4 competition, many of the top-performing entries were not single methods but carefully crafted ensembles, and machine-learning approaches featured prominently among the leaders (Makridakis et al., 2018). The subsequent M5 competition, focused on hierarchical retail sales data, extended these findings, again underscoring the importance of hybridization and the need for systematic, transparent evaluation (Makridakis et al., 2022). These competitions demonstrate not only that combinations can outperform their constituent models but also that they provide robustness across heterogeneous environments, which is critical in volatile financial and operational contexts. Equally vital is the rigor with which comparative testing is conducted. Differences in forecast performance should not be judged by raw error statistics alone but evaluated through statistically principled procedures such as the Diebold–Mariano test, which formally assesses predictive accuracy across competing models (Diebold & Mariano, 1995). Complementing this, proper scoring rules such as the continuous ranked probability score provide incentives for well-calibrated probabilistic forecasts rather than merely accurate point estimates. Taken together, these strands of evidence motivate both the design of explicit hybrid architectures and the adoption of multi-metric, statistically grounded evaluation protocols in modern model-development studies.

Review of Applied Science and Technology
Volume 04, Issue 01 (2025)
Page No: 01 – 34
Doi: 10.63125/nr1j8527

Domain-specific evidence from finance further substantiates the rationale for each component of the proposed hybrid architecture, highlighting both predictive efficacy and methodological depth.

Figure 1: Proposed Hybrid-LPX v1.0 architecture integrating Prophet (structural additive model)



Within the deep-learning tradition, long short-term memory (LSTM) architectures have exhibited strong performance in stock-market applications, particularly in forecasting directional movements and extracting profitable cross-sectional signals from high-dimensional financial data (Fischer & Krauss, 2018; Krauss et al., 2017). Surveys further illuminate the expanding footprint of deep learning in financial time-series forecasting, documenting not only its empirical successes but also the technical challenges that remain in areas such as overfitting, interpretability, and regime adaptation (Sezer et al., 2020). Parallel to this, Prophet's additive decomposition framework has demonstrated resilience across large-scale forecasting exercises, with empirical validations underscoring its aptitude for handling structural breaks, non-linear seasonalities, and calendar-driven irregularities (Patton et al., 2019; Taylor, 2019). In the probabilistic forecasting domain, autoregressive recurrent approaches such as DeepAR have advanced the practice of generating full-distribution predictions, which serve as essential inputs to risk-sensitive functionals including Value-at-Risk (VaR) and Expected Shortfall (ES) (Salinas et al., 2020). Beyond machine learning, classical statistical innovations remain pivotal. Quantile regression offers a direct, distribution-free route to VaR estimation by focusing explicitly on conditional quantiles (Koenker & Bassett, 1978). Complementing this, recent theoretical advances establish that ES, long regarded as a coherent but elusive risk measure, is jointly elicitable with VaR, thereby enabling unified estimation and principled evaluation frameworks (Fissler & Ziegel, 2016). These insights have been operationalized through dynamic semiparametric models that align with the empirical features of financial time series, including time-varying volatility and heavy tails (Patton et al., 2019). Finally, rigorous backtesting remains indispensable for evaluating tail-focused forecasts. Frameworks such as Kupiec's proportion-of-failures test provide benchmarks for unconditional coverage, while Christoffersen's conditional coverage test extends this scrutiny to independence properties of violations, jointly ensuring calibration and temporal robustness (Christoffersen, 1998; Kupiec, 1995). Collectively, these findings underscore not only the feasibility but also the necessity of integrating LSTM, Prophet, and distributional models within a hybrid architecture calibrated for financial forecasting and risk prediction.

The rationale for specifically integrating LSTM, Prophet, and XGBoost rests on their methodological complementarity, each addressing distinctive dimensions of financial time-series complexity. Prophet offers an additive decomposition framework that encodes structured priors for trend, seasonality, and holiday effects, augmented by automatic change-point detection. This design

closely mirrors equity market regularities, such as day-of-week anomalies, month-end effects, and earnings-cycle dynamics, making it particularly well suited for calendar-driven behaviors . By contrast, LSTM networks excel in capturing nonlinearities and state-dependent dynamics inherent in asset returns and realized volatility, where memory of past states and volatility clustering significantly influence future trajectories. Complementing both, XGBoost introduces sparse-aware gradient boosting that can flexibly learn interaction-rich relationships across heterogeneous predictor sets, encompassing technical indicators, cross-asset signals, and macroeconomic factors, while its efficient training and regularization mechanisms facilitate broad hyperparameter exploration and robust generalization (Chen & Guestrin, 2016). The intellectual lineage of gradient boosting, firmly grounded in additive regression trees (Friedman, 2001), links this approach to the wider ensemble literature, while random forests remain an important nonparametric comparator (Breiman, 2001). Within empirical finance, recent studies document that machine-learning predictors, especially those exploiting interactions and nonlinear structures, consistently outperform linear benchmarks, underscoring the relevance of such algorithms to asset-pricing and forecasting problems (Gu, Kelly, & Xiu, 2020). Parallel comparative investigations of Prophet across environmental and hydrological forecasting domains further attest to its robustness in handling complex or nested seasonalities, offering useful analogs for financial time series shaped by multiple trading calendars (Taylor & Letham, 2018). In a hybrid architecture, these distinct inductive biases can be synthesized through strategies such as stacking, where a meta-learner integrates their outputs, or through Bayesian model averaging, where probabilistic weighting reflects relative performance. Both approaches benefit from careful validation on rolling out-of-sample blocks, ensuring that the resulting ensemble captures path dependence, seasonal regularities, and cross-feature interactions while maintaining robustness across regimes.

The objective of this study is to design, formalize, and rigorously evaluate a hybrid forecasting framework that integrates three complementary learners an LSTM for sequence dynamics, Prophet for trend–seasonality with structural breaks, and XGBoost for interaction-rich tabular signals to enhance stock-price prediction and risk estimation in internationally diverse equity markets. Concretely, the work pursues eight measurable goals. First, specify an end-to-end architecture with clean interfaces among data ingestion, feature engineering, base-model training, and a hybrid combiner that produces both point and distributional forecasts. Second, implement three principled integration mechanisms stacking with a level-1 meta-learner, residual error-correction where a secondary learner models base-model residuals, and dynamic weighting with non-negative, time-varying weights so that the contribution of each base model can be quantified and contrasted. Third, operationalize a risk module that maps forecasts into volatility, Value-at-Risk, and Expected Shortfall by training quantile-aware components and evaluating their calibration with standard coverage diagnostics. Fourth, define a leakage-safe, walk-forward, nested cross-validation protocol that separates hyperparameter tuning, meta-learning, and final testing, enabling unbiased comparisons across baselines and hybrid variants. Fifth, establish an evaluation battery that covers point error (RMSE, MAE, sMAPE), directional accuracy (including class-imbalance-aware metrics), and distributional quality (quantile loss, CRPS), alongside formal comparative tests to determine whether observed gains are statistically reliable. Sixth, conduct ablation and sensitivity analyses to identify the marginal value of each base learner, each integration strategy, and key feature families across market regimes (bull, bear, high-volatility) and horizons (one-day and multi-day). Seventh, embed explainability throughout the pipeline global and local attributions for tree-based learners, weight-trajectory analysis for the hybrid, and interpretable decomposition artifacts so that improvements are traceable to model structure and inputs. Eighth, benchmark against strong, transparent baselines (random-walk, ARIMA/SARIMA, GARCH, and each standalone learner) and document a reproducible development path with deterministic training seeds, version-locked environments, data preprocessing scripts, and model cards that articulate scope and limitations. Collectively, these objectives align the paper with a development-first contribution: not merely reporting accuracy numbers, but presenting a validated, interpretable, and portable hybrid design ready to be implemented and stress-tested across assets and geographies in the accompanying implementation study.

LITERATURE REVIEW

The literature on financial time-series forecasting has evolved from classical econometrics to a mature, model-development ecosystem that blends deep sequence learning, structural time-series

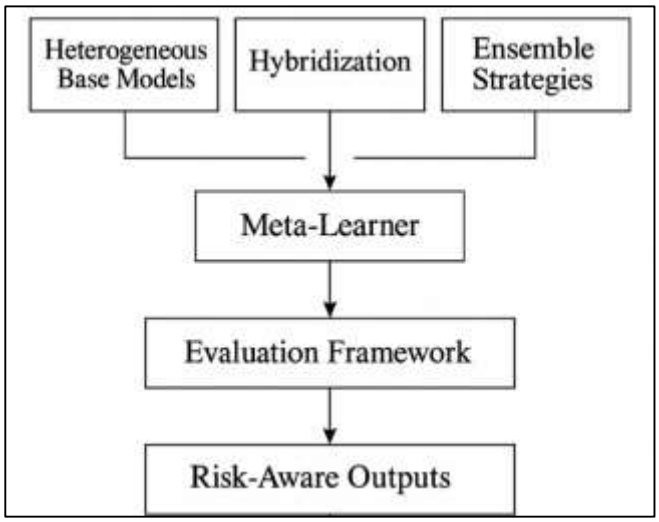
components, and gradient-boosted decision trees. Early strands established volatility dynamics and baseline predictability, while subsequent machine-learning research introduced flexible nonlinear function approximators and feature-rich pipelines. Within this landscape, three families of models consistently recur as strong yet complementary building blocks: long short-term memory (LSTM) networks that encode temporal dependence and nonlinear state transitions; additive models such as Prophet that represent trend, multiple seasonalities, holiday effects, and changepoints with interpretable components; and boosted tree ensembles such as XGBoost that learn interaction-dense mappings over heterogeneous technical, cross-asset, and macro features. Recent work emphasizes hybridization as a disciplined development strategy rather than an ad-hoc blend, motivating architectures that combine these learners through stacking with a meta-model, residual error-correction in which one learner models another's errors, and dynamic weighting schemes that adapt contributions across regimes. Parallel advances in probabilistic forecasting have shifted evaluation from point accuracy alone toward calibrated distributions, enabling direct estimation of risk functionals such as Value-at-Risk and Expected Shortfall via quantile-aware losses and proper scoring rules. This, in turn, has heightened the importance of leakage-safe validation (rolling origin, nested cross-validation), transparent baseline comparisons (random walk, ARIMA/SARIMA, standalone deep or tree models, and GARCH-family risk baselines), and formal significance testing to substantiate gains. The literature also underscores explainability and governance as first-class development requirements: decomposition plots from structural models, Shapley-value attributions for tree ensembles, and weight-trajectory diagnostics for hybrids help connect performance to mechanisms. Collectively, these strands point to a coherent agenda for model development: architect explicit information flows among LSTM, Prophet, and XGBoost; train and select hybrid variants under walk-forward protocols; quantify contributions through ablations and regime analyses; and report both predictive accuracy and risk calibration with decision-relevant metrics. Positioned this way, the review does not merely summarize prior results; it distills a design space and a methodology for constructing robust hybrid forecasters that are interpretable, reproducible, and suitable for international equity markets where calendar effects, structural breaks, and nonlinear dynamics coexist.

Hybrid and Ensemble Forecasting in Finance

The model-development literature on forecast combination provides a rigorous and principled foundation for constructing hybrid forecasters in finance, offering both theoretical justification and practical guidance. Central to this perspective is the recognition that each individual model represents a noisy, potentially biased lens on an underlying data-generating process. By exploiting diversity among these learners, combination strategies seek weighted schemes that minimize generalization error, thereby improving predictive reliability beyond what any single constituent model can achieve. Classical results demonstrate that, under mild assumptions, linear or nonlinear combinations can strictly dominate the mean-squared error of individual forecasts, particularly when base-model errors are partially uncorrelated and the constituent algorithms possess complementary inductive biases (Timmermann, 2006). Extensive reviews of decades of empirical practice reinforce this principle, concluding that pooling forecasts across heterogeneous specifications mitigates sensitivity to model misspecification, structural breaks, and regime shifts, yielding more robust and stable predictions (De Menezes et al., 2000). Beyond naïve averaging, contemporary frameworks formalize learning-to-combine as an explicit estimation problem: combination weights or meta-models are trained on historical data, with performance improvements arising from the interplay between variance reduction, bias management, and diversity among base learners (Yang, 2004). In macro-financial and high-frequency contexts, conditional and state-dependent combination strategies further refine this approach by dynamically adapting weights in response to evolving market conditions, acknowledging that no single model remains consistently optimal across different regimes (Aiolfi & Timmermann, 2006). Together, these strands provide a clear development agenda for hybrid forecasters in financial applications. The process begins with specifying a family of heterogeneous base models chosen for complementary strengths, proceeds through the design of a statistically valid protocol for learning combination weights or meta-predictions, and concludes with rigorous evaluation on genuinely out-of-sample data. This methodology not only enhances predictive performance but also supports robust risk assessment and governance, establishing a sound foundation for hybrid model deployment in complex financial environments.

Review of Applied Science and Technology
Volume 04, Issue 01 (2025)
Page No: 01 – 34
Doi: 10.63125/nr1j8527

Figure 2: Conceptual Framework of Hybrid and Ensemble Forecasting in Finance



Hybridization in time-series forecasting extends well beyond simple linear pooling by explicitly designing the flow of information among models with heterogeneous structural assumptions, and financial applications have provided a rich testbed for such approaches. A foundational stream of research combines linear stochastic components with nonlinear learners to simultaneously capture persistent trends, seasonalities, and complex higher-order interactions. Seminal examples include hybrid ARIMA–neural network architectures, in which residuals from a parsimonious linear model are passed as inputs to a nonlinear network that learns the remaining structure, thereby decomposing predictable linear behavior from more intricate dynamics (Zhang, 2003). In equity prediction, analogous constructions pair differenced or baseline linear models with margin-based learners, as in hybrid ARIMA–support vector machine pipelines, where ARIMA residuals are transformed into features that the nonlinear module uses for correction (Pai & Lin, 2005). These designs are explicitly developmental: they define interfaces for decomposition, residual modeling, and recombination, allowing researchers to evaluate the marginal contributions of each module and to identify potential failure modes. Subsequent work refined these templates through enhanced training regimes and systematic error allocation, demonstrating that hybrid ANN–ARIMA schemes can outperform either component in isolation while maintaining interpretability and computational tractability (Khashei & Bijari, 2011). Parallel lines of innovation explored multiresolution hybrids, particularly suited to noisy financial series, in which scale-localized signals are first extracted via wavelet transforms and then mapped nonlinearly using ensembles an approach naturally complementary to calendar-aware structural models and sequence learners such as LSTM (Wang et al., 2012). Across these studies, a common principle emerges: hybrid development begins by articulating the distinct role of each module, specifying training objectives aligned with the forecasting target, and integrating components under validation protocols that are robust to information leakage. By explicitly combining decomposition, sequence modeling, and interaction-rich tabular learning, these frameworks not only improve predictive performance but also facilitate interpretability, modular analysis, and governance, establishing a systematic template for the construction of hybrid forecasters in complex, real-world financial environments.

The ensemble perspective in time-series forecasting extends beyond static averaging to include dynamic pooling and structure-aware aggregation, both of which are especially pertinent in financial markets characterized by frequent regime transitions. Dynamic model averaging treats the weights assigned to competing predictors as latent, time-varying states that are updated continuously as new information becomes available. This allows the ensemble to adapt naturally when predictive relationships shift, offering a valuable mechanism for forecasting risk and returns across macro-financial cycles and structural breaks (Koop & Korobilis, 2012). Complementing this, research on neural-network ensembles has introduced operators and training heuristics designed to maximize diversity and stability among constituent models. These strategies produce performance gains that remain robust across forecast horizons and demand patterns, a property that directly

translates to financial applications where feature sets include a mix of technical indicators, cross-asset signals, and event-time effects (Kourentzes et al., 2014). From a development perspective, these insights provide a practical blueprint for hybrid LSTM–Prophet–XGBoost architectures. Structural components, such as Prophet, encode calendar effects, trends, and level shifts, while sequence learners like LSTM capture nonlinear temporal dependencies and volatility clustering. Boosted trees, implemented via XGBoost, can serve either as residual correctors or as meta-learners integrating base predictions with exogenous signals, exploiting interaction-rich relationships among heterogeneous features. Crucially, the combination-learning process must be framed as a supervised estimation problem, using out-of-fold predictions to prevent information leakage. Rigorous evaluation under rolling, regime-segmented validation further ensures stability, calibration, and robustness to shifts in market dynamics (Yang, 2004). Taken together, these studies underscore that hybrid and ensemble forecasting in finance is not a loose aggregation of techniques but a coherent methodology: specify complementary models, implement principled combination rules, and document predictive performance with transparent ablation studies, diagnostic checks, and backtests. By doing so, practitioners can build ensembles that are both interpretable and resilient, supporting decision-making in complex, dynamic financial environments.

LSTM and Deep Sequence Models for Equity Prediction

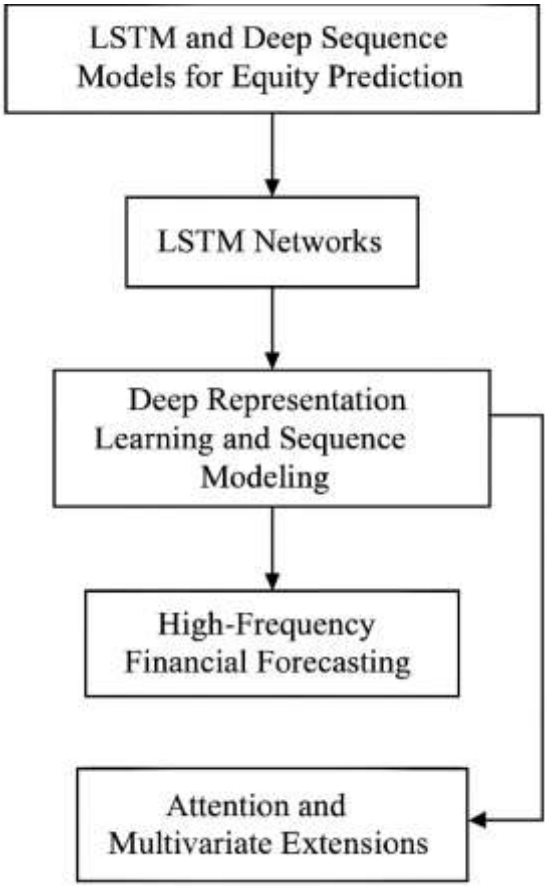
Long short-term memory (LSTM) networks have become a central pillar in equity forecasting because their gating mechanisms explicitly regulate the flow of information through time, enabling them to model nonlinear dependencies and long-range temporal structure in price and volume series. Early finance-focused studies demonstrated that LSTMs can outperform traditional baselines by learning from raw or lightly engineered technical indicators. For example, models trained on price histories augmented with momentum and volatility features achieved systematic directional accuracy advantages over conventional classifiers and shallow neural nets, indicating that LSTMs can internalize stylized facts such as volatility clustering and regime persistence (Ara et al., 2022; Nelson et al., 2017). Hybrid pipelines that pair deep representation learning with sequence modeling strengthened this pattern. A stacked autoencoder plus LSTM framework, trained end-to-end on market microstructure signals and daily bars, captured both denoised latent structure and dynamic evolution, yielding improved point forecasts and reduced reconstruction error relative to single-stage networks (Bao et al., 2017; Jahid, 2022). Beyond purely numerical inputs, multimodal designs that fuse quantitative and textual signals (e.g., news embeddings with OHLCV tensors) used an LSTM backbone to encode delayed, decaying effects of events on prices and to learn cross-sections of firm-specific sensitivities an architecture that aligns closely with equity analysts' practice of linking narrative catalysts to subsequent returns (Akita et al., 2016; Uddin et al., 2022). To better represent coexisting short-cycle and long-cycle forces in equities (e.g., intraday order-flow bursts overlaying weekly seasonality), convolution-recurrent hybrids such as LSTNet combine CNNs for local pattern extraction with recurrent blocks for long-horizon dependencies, and an explicit autoregressive head to stabilize scale sensitivity in nonstationary financial series (Lai et al., 2018; Akter & Ahad, 2022). Collectively, these developments position LSTM-centered architectures as a natural foundation for model development work that seeks to integrate heterogeneous signals into a unified, trainable forecasting engine.

As model development progressed into higher-frequency domains and richer data modalities, deep sequence models expanded their scope from daily bars to limit order books (LOBs) and intraday event streams. Leveraging billions of quotes and trades, deep-learning architectures revealed consistent mappings from order-flow histories to subsequent price movements, challenging the assumption that high-frequency dynamics are entirely idiosyncratic and demonstrating that data-driven representations can generalize across instruments, time intervals, and market conditions (Arifur & Noor, 2022; Sirignano & Cont, 2019). Architectures designed specifically for LOB tensors, such as DeepLOB, employ stacked convolutional layers to capture spatial depth across multiple price levels before feeding extracted representations into LSTM layers that track temporal evolution. This design produces stable out-of-sample classification accuracy on exchange data and demonstrates transferability across previously unseen tickers (Rahaman, 2022; Zhang et al., 2019). Concurrently, research on intraday bar data has shown that ensemble strategies further enhance model robustness. By combining multiple LSTM networks trained on diverse technical indicator sets and dynamically reweighting them according to recent predictive performance, these ensembles adapt to nonstationarities and micro-regime shifts that are characteristic of intraday equity markets

Doi: 10.63125/nr1j8527

(Borovkova & Tsiamas, 2019; Hasan et al., 2022). Even convolutional architectures such as CNNpred, while not inherently recurrent, contribute complementary local-pattern features extracted from multi-market and macroeconomic proxies. When integrated with sequence modules, these features improve generalization by disentangling cross-sectional co-movements from idiosyncratic dynamics within individual series (Hoseinzade & Haratizadeh, 2019; Hossen & Atiqur, 2022). Taken together, these strands inform a coherent model-development agenda for high-frequency financial forecasting. LSTM layers serve as the temporal core, while surrounding convolutional feature extractors and ensemble logic stabilize training, mitigate the effects of regime changes, and capture hierarchical, multi-resolution structures in equity signals. This combination enables robust predictions, supports transferability across instruments, and provides a foundation for designing hybrid architectures that can operate effectively in complex, high-frequency trading environments.

Figure 3: Theoretical Framework of LSTM for Equity Prediction



A key practical challenge in developing LSTM-centric equity models lies in balancing model capacity with interpretability and calibration across a multitude of correlated inputs, including prices, volumes, technical factors, macroeconomic indicators, and textual signals. Attention-augmented recurrent neural networks provide one pathway to address this challenge. Dual-stage attention mechanisms rank exogenous drivers at each time step while simultaneously selecting relevant past hidden states, enabling the network to highlight which features and temporal lags are most influential for a given prediction. This approach not only improves transparency for model diagnostics but also aligns the system with domain knowledge regarding market catalysts and decay patterns (Qin et al., 2017). Another axis of model design involves multivariate sequence modeling at scale. Multivariate LSTM-FCN architectures, which combine temporal recurrence with fully convolutional blocks and channel-wise recalibration, efficiently handle high-dimensional series while preserving discriminative temporal patterns. These properties are particularly valuable in equity universes where features share periodic rhythms but vary in amplitude and noise characteristics (Karim et al., 2019).

Such architectures facilitate feature compression, temporal abstraction, and robust encoding of both short- and long-horizon dependencies, making them suitable for integration into hybrid forecasting stacks. Taken together, the literature suggests a coherent blueprint for robust equity forecasting: first, employ convolutional or autoencoder modules to compress high-granularity, noisy inputs; second, route the compressed sequences through LSTM or attention-enhanced LSTM layers to encode temporal dependencies across multiple horizons; and third, incorporate ensemble strategies or calibration heads to stabilize predictions across market regimes and enhance out-of-sample reliability. This framework directly motivates the present study's hybrid development focus, wherein LSTM-based temporal encoding is designed to interoperate seamlessly with complementary modules, including tree-boosted residual learners and trend/seasonality components, forming a single, testable forecasting stack capable of producing both accurate point forecasts and well-calibrated predictive distributions.

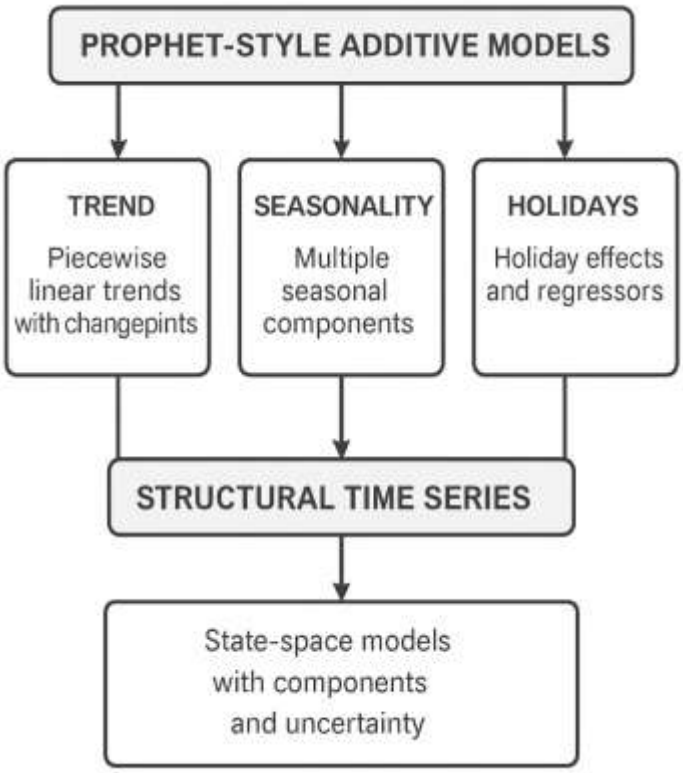
Prophet-Style Additive Models for Financial Forecasting

Additive, component-based time-series models provide a rigorous foundation for capturing calendar effects, regime shifts, and trend reconfigurations that are prevalent in equity markets. In state-space formulations of exponential smoothing, a broad family of additive trend and seasonality models emerges, offering likelihood-based estimation, automatic model selection, and probabilistic forecasts, which make such models attractive building blocks within hybrid pipelines by supplying interpretable, calibrated, and structured representations of underlying series dynamics (Hyndman et al., 2002; Tawfiqul et al., 2022). For series exhibiting multiple overlapping seasonalities, analogous to intra-week trading rhythms layered on annual reporting calendars, TBATS extends exponential smoothing by incorporating Box-Cox transformations, ARMA error terms, and trigonometric Fourier components, delivering a scalable approach to modeling complex seasonal patterns and damped trends frequently observed in market activity measures (De Livera et al., 2011; Kamrul & Omar, 2022). Earlier operational research in electricity demand forecasting demonstrated how double and triple seasonal exponential smoothing can effectively accommodate interacting daily and weekly cycles, and this methodology transfers naturally to equity markets, where trading calendars, month-end effects, and holidays interact in predictable ways (Mubashir & Abdul, 2022; Taylor, 2003). Complementary regression-with-ARIMA-errors approaches, widely used in official statistics, model trading-day and holiday effects as explicit regressors while allowing serially correlated residuals, producing transparent decomposition artifacts suitable for practitioner audit and governance (Findley et al., 1998; Reduanul & Shoeb, 2022). Together, these strands support the inclusion of an additive, interpretable layer in hybrid forecasters, whereby trend, seasonality, and holiday structure are encoded in a dedicated additive component, adjusted residuals are passed to nonlinear learners, and additive forecasts may serve as meta-features in stacking with sequence and tree-based modules, positioning the additive block as a stabilizer that reduces information leakage through fixed calendar regressors and deterministic Fourier bases and provides governance-ready diagnostics including component plots and revision measures, which are essential for finance, thereby contributing both predictive stability and transparency while complementing the nonlinear, interaction-driven elements of modern hybrid architectures.

Bayesian structural time series (BSTS) generalizes the additive modeling paradigm by embedding components such as local level and slope, seasonal cycles, and regression effects within a probabilistic state-space framework that provides explicit uncertainty quantification and dynamic variable selection, making it particularly suitable for financial forecasting. For nowcasting and "predicting the present," BSTS integrates contemporaneous high-dimensional signals, including search trends and event proxies, through spike-and-slab priors, producing sparse and interpretable regression terms layered on top of structural components (Sazzad & Islam, 2022; Scott & Varian, 2014). The framework also supports counterfactual forecasting for policy or event evaluation via diffusion-regression state-space models, allowing market researchers to quantify the impact of interventions such as index reconstitutions or trading halts while maintaining additive trend and seasonality structure (Brodersen et al., 2015; Sazzad & Islam, 2022). A central consideration in additive modeling for finance is the robust handling of structural breaks and changepoints, which frequently occur at earnings cycles, macro announcements, or regime pivots. Multiple-break econometrics provides practical tools for locating and estimating such changes (Bai & Perron, 2003), while modern changepoint detection offers linear-time algorithms capable of scaling to thousands of securities and long histories (Killick et al., 2012; Noor & Momena, 2022). From a model-development

perspective, these insights translate into design choices for an additive layer: include piecewise-linear trends with data-driven break selection, encode holiday and event regressors with shrinkage priors, and learn stochastic seasonal components that evolve over time. When residuals deviate from Gaussianity, as is common in returns, state-space likelihoods and filtering can be adapted to non-Gaussian families without abandoning the additive structure, preserving interpretability while improving calibration of predictive intervals under volatile regimes (Adar & Md, 2023; Durbin & Koopman, 1997). The resulting additive block is probabilistic, break-aware, and event-sensitive, providing a foundation within a hybrid pipeline to enhance both forecast accuracy and risk estimation.

Figure 4: Prophet-Style Additive Models Framework for Financial Forecasting



For hybrid model development, which is the central theme of this study, the additive layer fulfills three complementary roles: decomposer, forecaster, and meta-feature generator. First, as a decomposer, it isolates calendar effects and slow-moving levels, producing cleaner, approximately stationary residuals that can be passed to sequence learners such as LSTM networks, which specialize in capturing nonlinear dependencies over short- and medium-term horizons. Second, as a stand-alone forecaster, the additive layer generates base predictions whose error profiles differ systematically from those of deep sequence or gradient-boosted models, enhancing diversity for stacking or dynamic weighting. In multi-seasonal contexts, TBATS-style Fourier expansions provide compact and leakage-safe encodings that can be shared across assets, facilitating large-scale training without bespoke feature engineering (De Livera et al., 2011; Qibria & Hossen, 2023). When trading-day and holiday effects are material, regARIMA-style holiday regressors with ARIMA errors serve as transparent governance artifacts and can be reused across retraining cycles (Durbin & Koopman, 1997; Findley et al., 1998). Third, as a meta-feature generator, structural components including level, slope, seasonal indices, and break indicators become explanatory signals for tree-boosted meta-learners, allowing interpretable interactions between decomposed structural effects and technical or macro covariates. Within Bayesian structural time-series frameworks, sparsity-inducing priors naturally select among candidate event and calendar regressors, yielding stable and audit-friendly component sets that generalize across international equities with heterogeneous calendars (Istiaque et al., 2023;

Scott & Varian, 2014; Taylor, 2010). Finally, changepoint tools, including multiple-break estimation and fast detection algorithms, integrate into the retraining loop to refresh trend segments and maintain calibration through periods of heightened volatility, while non-Gaussian state-space variants preserve the reliability of predictive intervals used downstream for Value-at-Risk and Expected Shortfall estimation (Bai & Perron, 2003). In sum, Prophet-style additive modeling and its state-space extensions provide an interpretable, regime-aware backbone that a hybrid LSTM–Prophet–XGBoost architecture can leverage to enhance both point forecast accuracy and risk calibration in equity markets.

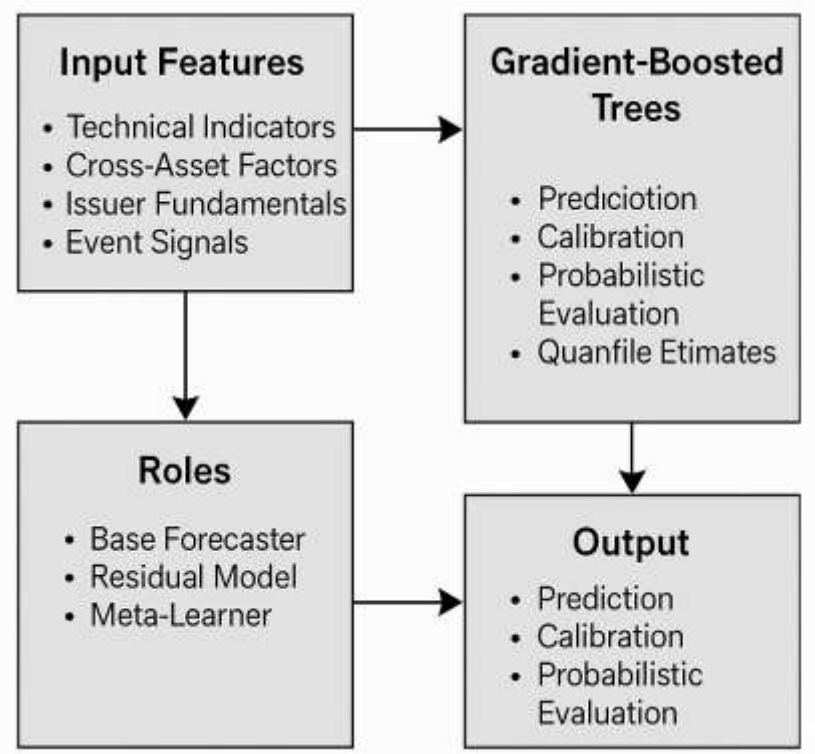
Gradient-Boosted Trees and XGBoost-Style Learners for Tabular Financial Time Series

Gradient-boosted decision trees have become a mainstay of model development for tabular financial forecasting because they combine strong nonlinear function approximation with pragmatic engineering features, including regularization, missing-value handling, monotonic constraints, and flexible loss functions, without requiring the extensive preprocessing often necessary for deep networks. Stochastic gradient boosting formalized iterative, shrinkage-based tree fitting with sub-sampling to control variance and improve generalization, establishing a robust template for scalable ensembles (Friedman, 2002; Akter, 2023). Modern implementations extend this framework with sparse-aware split finding, histogram binning, and parallelization, while widely used toolkits provide a broad menu of objectives such as squared error, logistic, Poisson, and Tweedie losses, enabling developers to tailor models to returns, directional classification, or counts of market events. The "xgboost" software family operationalizes regularized tree boosting with efficient solvers and distributed training, making gradient boosting practical at the scale of financial data warehouses and supporting disciplined hyperparameter search (Chen, He, Benesty, Khotilovich, & Tang, 2016). Concurrently, research on categorical-feature treatment and ordered boosting reduces prediction shift and target leakage risks in settings with high-cardinality identifiers such as issuer, sector, or venue, which are common in market microstructure and cross-sectional stock selection (Hasan et al., 2023; Prokhorenkova et al., 2018). Tutorials and comparative reviews now codify best practices, including trade-offs between learning rate and trees per iteration, depth versus leaf-wise growth, and early stopping, and document scenarios in which boosted trees outperform other tabular learners, including random forests and linear models with interactions (Masud et al., 2023; Natekin & Knoll, 2013). For model-development pipelines intended for hybridization with sequence learners and structural components, these gradient-boosted tree capabilities provide a stable, interpretable, and computationally efficient backbone for tabular features such as technical indicators, cross-asset factors, issuer fundamentals, and event-related signals.

A second attraction of tree boosting for financial applications is its compatibility with distributional and risk-aware objectives. While boosted trees are often applied to point prediction, related random forests and ensemble variants can estimate conditional quantiles and full conditional distributions, enabling the computation of Value-at-Risk, Expected Shortfall, and prediction intervals that can be evaluated using coverage diagnostics (Sultan et al., 2023; Meinshausen, 2006). Generalized random forests extend this capability beyond means and quantiles to any parameter identified by local moment conditions, which is useful for analyzing heterogeneous treatment effects around events or regime-dependent elasticities in factor models (Athey et al., 2019; Hossen et al., 2023). In practice, converting boosted-tree scores into reliable risk measures requires probability or quantile calibration, as classic studies have shown that raw outputs from margin-based and tree ensembles can be miscalibrated, while post-hoc calibrators such as Platt scaling and isotonic regression substantially improve reliability (Tawfiqul, 2023; Zadrozny & Elkan, 2002). From a model-development perspective, these insights inform three concrete design choices for a hybrid architecture: first, adopt quantile or asymmetric loss variants during training of boosted components to align directly with VaR and tail-risk objectives; second, incorporate calibration on out-of-fold predictions to prevent leakage; and third, persist calibration artifacts, including calibration curves and reliability diagrams, alongside the model for governance purposes (Shamima et al., 2023). Because boosted trees natively compute feature importance and support SHAP-style attributions, developers can trace risk forecasts to specific drivers, including macroeconomic surprises, liquidity proxies, or structural features originating from an additive layer, supporting an explainability-first approach that is essential for documentation, auditing, and model-risk management in financial applications (Ashraf & Ara, 2023).

Review of Applied Science and Technology
Volume 04, Issue 01 (2025)
Page No: 01 – 34
Doi: 10.63125/nr1j8527

Figure 5: Gradient-Boosted Trees and XGBoost-Style Learners for Tabular Financial Time Series



Empirical applications in finance strongly support these design choices across diverse use cases, demonstrating how carefully engineered hybrid pipelines can leverage ensemble learning to improve predictive performance and risk assessment. In credit risk, large-scale benchmark studies indicate that modern ensembles, including gradient-boosted decision trees, frequently outperform alternative approaches across realistic scenarios characterized by class imbalance and cost-sensitive misclassification, provided that input features are systematically constructed and validation protocols are rigorously leakage-free (Lessmann et al., 2015; Sanjai et al., 2023). In equity prediction, boosted trees excel as meta-learners that integrate heterogeneous base forecasts and exogenous signals, or as residual-correcting models that capture the information left unexplained after structural additive components remove calendar effects and trend dynamics. In these configurations, gradient boosting exploits its strength in modeling nonlinear interactions among lagged returns, cross-asset spreads, technical indicators, and regime-dependent flags, providing interpretable yet flexible mappings that complement sequence and structural layers. The maturity of supporting software further enhances practical adoption, with R and Python implementations of XGBoost offering reproducible, production-ready pipelines, deterministic seeds, monotonicity constraints to enforce economically sensible factor relationships, and GPU acceleration to enable rapid walk-forward tuning and hyperparameter optimization (Bentéjac et al., 2021; Chen et al., 2016; Akter et al., 2023). Within the hybrid architecture central to this study, the boosted-tree component is explicitly designed to fulfill three roles: first, as a base forecaster trained on tabular inputs to deliver a diversely biased prediction stream; second, as a residual model that learns the systematic errors of LSTM and additive layers; and third, as a level-1 meta-learner that stacks base predictions with macroeconomic and market covariates under strictly out-of-fold evaluation protocols (Razzak et al., 2024). Across all roles, outputs from the boosted-tree models will be paired with calibration procedures, probabilistic assessment, and quantile evaluation to ensure that gains in point accuracy translate into improved risk estimation, robust predictive distributions, and decision-relevant performance metrics that are critical for both research and applied financial forecasting(Niculescu-Mizil & Caruana, 2005).

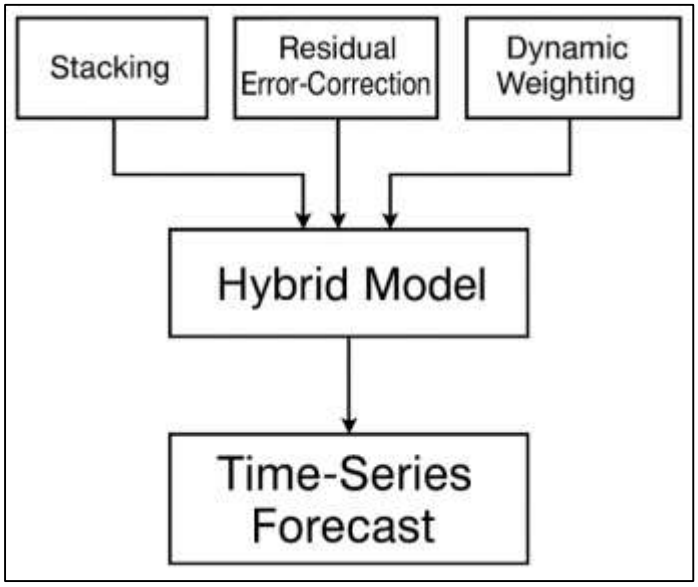
Stacking, Residual Error-Correction, and Dynamic Weighting for Time-Series Hybrids

Hybridization in financial forecasting achieves its greatest effectiveness when designed as a learnable, principled system rather than an ad-hoc mixture of models. Three main families of design dominate modern model development: stacking, residual error correction, and dynamic weighting.

Review of Applied Science and Technology
Volume 04, Issue 01 (2025)
Page No: 01 – 34
Doi: 10.63125/nr1j8527

Stacking implements the "learn to combine" principle by training a meta-learner on out-of-fold predictions generated by base learners such as LSTM, Prophet, and XGBoost under leakage-free splits, producing a meta-model that optimally minimizes generalization error (Breiman, 1996; Ting & Witten, 1999). The Super Learner framework extends this concept by searching over a library of potential meta-combiners and selecting an asymptotically optimal convex combination under cross-validation, a property particularly appealing to practitioners who require defensible, repeatable ensemble designs (Istiaque et al., 2024; Laan et al., 2007). Complementing this, classical combination theory generalized simple averaging to regression-based pooling that conditions on the levels of base forecasts, yielding interpretable weights that reflect the relative contribution of each learner (Granger & Ramanathan, 1984). Bayesian Model Averaging offers a probabilistic alternative by explicitly accounting for model uncertainty and generating calibrated predictive distributions, which is particularly useful when hybrid ensembles are deployed for risk-sensitive financial applications (Hoeting et al., 1999; Akter & Shaiful, 2024). More recent innovations in feature-based model averaging demonstrate how meta-learners can map time-series characteristics, such as trend magnitude, seasonality strength, and spectral peaks, to weights over candidate models, operationalizing stacking at scale while providing interpretable drivers for ensemble decisions (Hasan et al., 2024; Montero-Manso & Hyndman, 2020). Collectively, these approaches reframe ensembling as a supervised learning problem with explicit objectives, constraints, and diagnostics, replacing heuristic averaging with a systematic methodology that is well-suited for hybrid LSTM–Prophet–XGBoost architectures applied across large portfolios of equities.

Figure 6: Stacking, Residual Error-Correction, and Dynamic Weighting for Time-Series Hybrids



Residual error correction complements stacking by explicitly directing information that a base model fails to capture into a second learner trained on the base model's residuals. In practical implementation, one can first fit a structural additive model, such as a Prophet-style or state-space formulation, to encode trends, multiple seasonalities, and holiday effects, and then train a nonlinear learner, for example gradient-boosted trees, on its residuals to capture interactions, thresholds, and regime-dependent behaviors. This "explain the remainder" approach generalizes early hybrid ARIMA–neural network strategies by clarifying interfaces, specifying objective functions, and enforcing leakage-aware resampling at each stage (Khashei & Bijari, 2012; Tawfiqul et al., 2024). Hybridization with recurrent components has been validated at scale; in the M4 forecasting competition, a combined exponential-smoothing and recurrent neural network architecture produced both accurate and stable forecasts, demonstrating how a structural block can stabilize levels and seasonality while a sequence learner models residual dynamics (Smyl, 2020; Subrato & Md, 2024). From a model-development perspective, residual correction encourages modular design, with each block trained for a distinct function, and supports compatibility with probabilistic

or quantile losses, enabling residual learners to target tail behavior. Ablation studies further quantify the marginal contribution of each module. Beyond two-stage designs, "stacked residuals" generalize the concept by iteratively fitting learners to the errors of the current ensemble, operating across heterogeneous model classes in a manner akin to boosting(Ashiqur et al., 2025), while allowing principled early stopping, nested validation, and interpretable error decomposition. In fields adjacent to equities, online aggregation of specialized experts has demonstrated that residual-style correction combined with adaptive weighting can be framed as regret-minimization problems with provable performance bounds, reinforcing the value of learnable correction layers in nonstationary and dynamically evolving financial environments (Breiman, 1996; Devaine et al., 2013; Ting & Witten, 1999).

Dynamic weighting completes the hybridization framework by allowing combination weights to evolve in response to changing market regimes, time horizons, or feature conditions, which is essential when no single base learner remains dominant across all periods. While Bayesian Model Averaging provides a probabilistic approach to averaging under model uncertainty, practitioners frequently prefer data-driven schemes that adjust weights based on recent forecast performance or meta-features that proxy for predictability conditions, with empirical evidence from energy-price forecasting demonstrating that such adaptive pooling improves stability and out-of-sample accuracy (Bordignon, Candelon, & Sy, 2013). In supervised implementations, feature-weighted linear stacking directly maps meta-features to weights, enabling the combiner to emphasize base models in contexts where they excel, such as when LSTM captures volatility bursts, Prophet encodes calendar-driven effects, and XGBoost leverages rich exogenous signals (Hasan, 2025; Sill et al., 2009). Guardrails are equally important: dynamic pooling must be trained on out-of-fold histories to prevent look-ahead bias, and regularization or convexity constraints should be applied to avoid overreacting to short-term noise (Granger & Ramanathan, 1984; Sultan et al., 2025). Operational pipelines should integrate transparent diagnostics, including partial-dependence plots for meta-features, weight-trajectory visualizations, and rolling evaluations segmented by regime, so that shifts in ensemble behavior are traceable and governed. Taken together, stacking, residual error correction, and dynamic weighting constitute a coherent development methodology: each module has a clear role in handling temporal, structural, or exogenous information; combinations are learned on leakage-free out-of-fold predictions; residual learners capture structured misspecification; and weights adapt through meta-features or recent loss histories, with calibration and interpretability diagnostics embedded throughout, ensuring that hybrid LSTM–Prophet–XGBoost architectures can deliver both accurate forecasts and reliable risk estimates in dynamic financial environments (Bordignon et al., 2013; Granger & Ramanathan, 1984; Sill et al., 2009).

Risk Forecasting (Volatility, VaR, and ES)

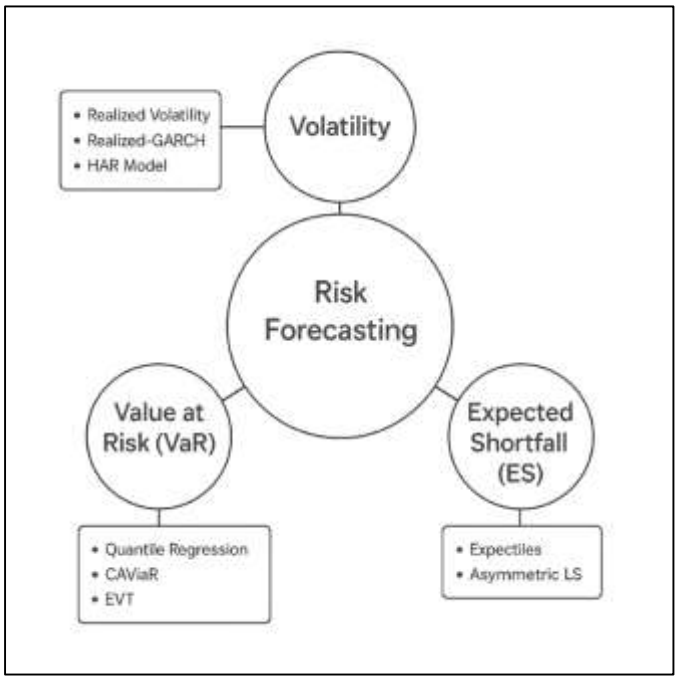
Volatility is the principal conduit through which models translate market dynamics into risk, and advances in measurement have transformed how volatility is forecast in model development. High-frequency–based estimators most notably realized volatility provide efficient, low-bias proxies for latent conditional variance and thereby offer richer signal content than daily return–only approaches (Andersen et al., 2003). Robust handling of market microstructure noise via realized kernels further stabilizes these measures, enabling practitioners to incorporate intraday information without inflating variance from bid–ask bounce or asynchronous trading (Barndorff-Nielsen et al., 2008). Parsimonious long-memory structures, such as the HAR-RV specification, then map multi-scale persistence into tractable dynamics that suit downstream learning modules (Corsi, 2009). Joint models like Realized-GARCH formalize the connection between realized measures and the conditional variance with a measurement equation, producing variance forecasts that align with the observed ex-post variation while retaining a likelihood-based inference pathway needed for disciplined model comparison (Hansen et al., 2012; Sanjai et al., 2025). Within a hybrid pipeline, these elements supply a curated feature stack daily RV, multi-horizon HAR lags, and kernel-based noise-resistant signals that can be ingested by gradient boosting or recurrent learners to improve volatility nowcasts/forecasts and stabilize the risk block that ultimately drives VaR and ES calculations.

Translating conditional variance information into risk metrics requires frameworks that target tail behavior directly. Quantile-dynamic approaches such as CAViaR update Value-at-Risk through autoregressive evolution of conditional quantiles, matching the object of interest without imposing a full distribution and thus offering a development path that couples neatly to ML-produced covariates or state variables (Engle & Manganelli, 2004). Extreme-value–augmented schemes, in

Review of Applied Science and Technology
Volume 04, Issue 01 (2025)
Page No: 01 – 34
Doi: 10.63125/nr1j8527

turn, estimate the tail of the innovation distribution and splice it into conditional scale dynamics, yielding VaR/ES estimators that are sensitive to rare events a desirable property when models are trained on rolling windows with regime shifts (McNeil & Frey, 2000). The coherence of Expected Shortfall provides a theoretically disciplined target for loss control and model calibration, especially when risk aggregation is required across desks or factors (Acerbi & Tasche, 2002). Expectile-based methods link ES and VaR to asymmetric least squares, creating a smooth objective that is convenient for machine-learning solvers and facilitates joint estimation or ensembling with quantile modules (Taylor, 2008). For development workflows, these families define loss functions and diagnostics that are aligned with the quantities regulators and internal risk policy care about, ensuring that enhancements from LSTM sequence modeling or tree-based learners translate into improvements in usable, well-defined risk outputs. Model evaluation and operationalization close the loop from volatility features to deployable risk numbers. Long/short asymmetries and tail skewness matter for risk limits, and empirical evidence shows VaR performance can depend on modeling skewed heavy-tailed errors and on recognizing that the short-side risk structure differs from the long-side (Giot & Laurent, 2003). Density-level evaluation helps diagnose whether a proposed risk engine delivers calibrated tail probabilities across time, complementing simple exception counts and strengthening governance for threshold setting and limit management (Berkowitz, 2001). In a hybrid architecture oriented toward development, these insights guide design choices: realized-measure inputs reduce proxy noise in the targets used to train volatility learners; quantile/expectile objectives align the ML loss with VaR/ES definitions; and EVT-style tail modeling, applied to scale-normalized residuals from LSTM/Prophet/XGBoost stacks, sharpens tail extrapolation without over-fitting the center of the distribution. The result is a risk-forecasting block that is modular so it can be stress-tested, backtested, and swapped component-wise yet integrated tightly enough to capture volatility clustering, multi-scale persistence, and tail risk within a single, auditable development framework (Acerbi & Tasche, 2002; Andersen et al., 2003; Barndorff-Nielsen et al., 2008).

Figure 7: Risk Forecasting Framework Integrating Volatility, Value at Risk (VaR), and Expected Shortfall (ES)



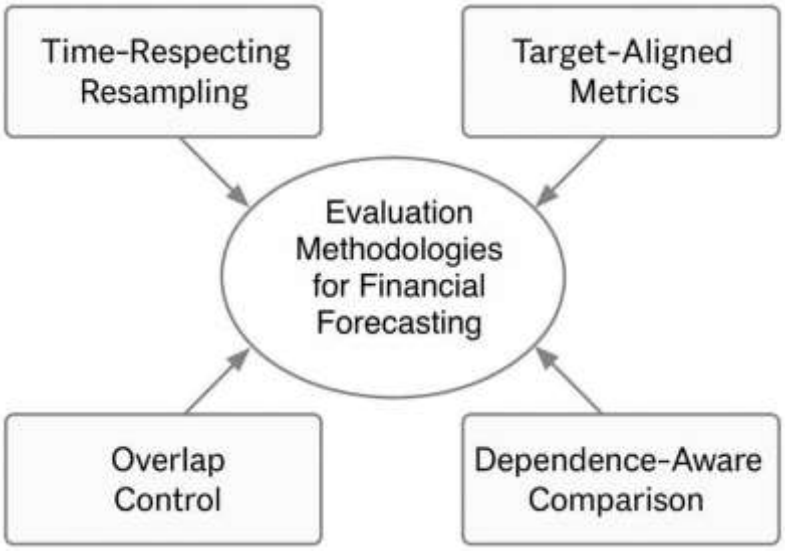
Evaluation Methodologies for Financial Forecasting

Reliable evaluation is the cornerstone that transforms model development into credible evidence, particularly for hybrid frameworks that integrate sequence learners, additive components, and gradient-boosted trees. Financial time series are inherently temporally dependent and often nonstationary, so the first guiding principle is to preserve the temporal order in every split. Rolling-origin evaluation, expanding-window validation, and walk-forward testing provide honest

Doi: 10.63125/nr1j8527

assessments of out-of-sample performance, while hyperparameter tuning and ensemble weight optimization must be nested within each training window to ensure that the test fold remains fully untouched and leakage is prevented. The second principle emphasizes alignment between evaluation metrics and forecasting targets. Point forecasts should be assessed using error magnitudes, such as mean absolute error, root-mean-squared error, or mean absolute scaled error, whereas directional or classification tasks benefit from metrics that account for class imbalance, including F1 score, Matthews correlation coefficient, and balanced accuracy. Probabilistic outputs, which are central for risk-oriented applications, should be evaluated using strictly proper scoring rules, including the logarithmic score and continuous ranked probability score, as well as quantile- and expectile-based losses to directly measure calibration for Value-at-Risk and Expected Shortfall. A third principle addresses target overlap: multi-step horizons can generate overlapping forecast targets, which inflates apparent sample size and induces serial dependence in residuals, so evaluation should employ non-overlapping anchors or embargo periods to mitigate these effects. Beyond these operational mechanics, evaluation must clearly separate selection from assessment(Arlot & Celisse, 2010). Hyperparameters, base-model weights, and meta-learner parameters are selected exclusively on validation folds, with final performance estimated only on a holdout or terminal rolling window. Comparative testing must also account for time-series dependence and model nesting. Conditional predictive ability tests, for instance, examine whether one model's forecast errors contain incremental information not captured by another, conditional on the information set available at forecast time, which is particularly relevant when hybrids are designed to exploit complementary signals (Giacomini & White, 2006). Cross-validation theory further informs procedure selection, as standard k-fold CV can be ill-posed under temporal dependence. Surveys and methodological reviews indicate that blocked CV, leave-future-out, and rolling schemes provide more reliable guidance for hyperparameter tuning, model selection, and error estimation in dependent series (Arlot & Celisse, 2010). Collectively, these protocols time-respecting resampling, target-aligned metrics, overlap control, and dependence-aware comparisons form the backbone of a defensible evaluation framework for hybrid financial forecasters, ensuring that improvements in point accuracy, probabilistic calibration, and risk measurement are credible, reproducible, and actionable in real-world settings.

Figure 8: Evaluation Methodologies for Financial Forecasting



When models are nested, such as when a hybrid meta-learner augments a strong baseline with additional features or base forecasts, standard equal-accuracy tests can be biased in favor of the larger model. Approximately normal tests that adjust for the extra noise introduced by estimating additional parameters provide a principled solution and are widely used to determine whether the augmented model genuinely improves predictive accuracy in nested settings (Clark & West, 2007). Volatility and risk forecasting require additional considerations: because financial return distributions

are often heavy-tailed and asymmetric, it is important to evaluate both unconditional and conditional calibration, including tail hit rates and the independence of exceedances, while pairing score-based assessments with backtests and reporting horizon-specific performance to avoid misleading averages. Comparative studies of volatility models demonstrate that even sophisticated specifications must be benchmarked against simple but resilient baselines to demonstrate practical superiority, a lesson that directly informs the use of modern machine-learning components as volatility or risk engines (Hansen & Lunde, 2005). Documentation is equally critical for evaluation: confidence intervals should be reported for performance differences, learning curves should be published, reliability diagrams for probabilistic outputs should be included, and ablations should quantify the marginal contribution of each model component, including LSTM alone, Prophet alone, XGBoost alone, stacked hybrids, and residual-corrected hybrids. For assessments oriented toward deployment, statistical improvements should be translated into decision-relevant metrics, such as turnover-adjusted returns, drawdown statistics, and tail-loss reductions, computed under realistic frictions. Taken together, dependence-aware tests for conditional and nested predictive ability, robust benchmarking norms for volatility and risk engines, and rigorous horizon-specific scoring ensure that evaluation is closely tied to the hybrid model's intended purpose, verifying that each added layer, whether sequence-based, structural, or tree-based, contributes reproducible and decision-relevant gains rather than fragile, selection-induced artifacts.

Explainability, Stability, and Regime Sensitivity

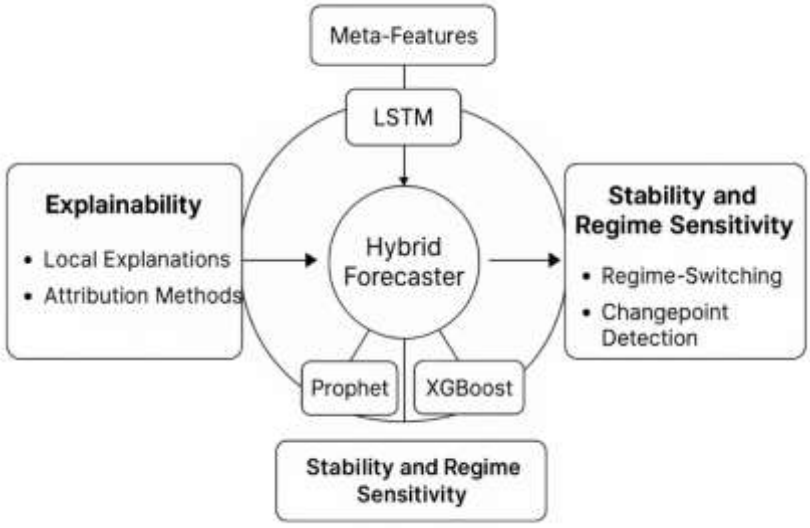
Building a hybrid forecaster that meaningfully integrates LSTM, Prophet, and XGBoost requires explainability that faithfully reflects each component's inductive bias while remaining stable enough to satisfy model governance requirements. Post-hoc local explanation tools provide a practical entry point for tracing predictions back to features and individual model components within a disciplined development workflow. Local surrogate methods approximate complex decision surfaces in the vicinity of a given prediction, producing human-readable attributions that can be aligned with domain intuition, such as a day-of-week effect captured by the additive layer or a volatility spike encoded in sequence memory. When embedded in an evaluation loop, these local explanations help diagnose situations where the hybrid relies on brittle signals, highlight interactions that suggest new engineered features, and reveal divergences between the drivers of point forecasts and those of tail-risk estimates (Hamilton, 1989). Path-integral or integrated-gradient attribution methods in deep networks complement surrogate approaches by enforcing axioms such as sensitivity and implementation invariance, which reduce ambiguity in credit assignment across timesteps and input channels, a particularly important consideration when LSTM layers encode overlapping horizons and correlated indicators. Explanation systems themselves must also be validated; stress tests that randomize model parameters or input data while holding architecture constant can expose methods that produce superficially plausible but uninformative attributions (Adebayo et al., 2018). In a rigorous development pipeline, explainability is treated as a first-class diagnostic, producing local explanations for representative market conditions, comparing them against additive-model components and boosted-tree splits, and gating model promotion based on qualitative consistency across neighboring points and quantitative stability of attribution rankings across refits. These practices mitigate the risk that apparent hybrid gains are driven by spurious or unstable features and strengthen the link between the model architecture and the mechanisms it claims to encode (Ribeiro et al., 2016).

Equally central to hybrid model development is the stability of forecasts under regime change and structural breaks, which is a critical consideration in equity markets where expansions and contractions produce markedly different volatility, liquidity, and microstructure characteristics. A forecaster that performs well only under a single regime risks generating unreliable point predictions and risk estimates, making downstream decision-making fragile. Regime-switching models formalize this challenge by allowing model parameters, latent states, and error structures to vary depending on the prevailing market regime. Outputs such as estimated state probabilities can be incorporated as meta-features that modulate ensemble weights, dynamically adjusting the influence of sequence learners, additive components, or boosted-tree modules in accordance with current market conditions. Complementing regime-switching, fast multiple-changepoint detection algorithms monitor residuals or structural components for shifts in trend or variance, triggering retraining, resetting state estimates, or reverting to simpler baselines when evidence of instability is detected. In practice, regime probabilities and break indicators are integrated into the hybrid feature interface:

Review of Applied Science and Technology
Volume 04, Issue 01 (2025)
Page No: 01 – 34
Doi: 10.63125/nr1j8527

the additive layer contributes trend, level, and seasonal components that are robust to calendar changes; the sequence layer, such as LSTM, operates on break-adjusted sequences to prevent the confounding of pre- and post-shift dynamics; and the boosted-tree layer captures conditional interactions that vary across regimes, including cross-asset and macro signals. Evaluation protocols mirror this design, requiring the test horizon to be segmented by inferred regimes and for accuracy, calibration, and tail coverage to be reported within each segment. This regime-aware evaluation ensures that performance claims are granular, auditable, and resilient to temporal heterogeneity. Beyond improving point and distributional accuracy, embedding regime detection and break handling tempers forecast variance over time, producing smoother weight trajectories, more stable SHAP or attribution profiles, and interpretable signals that facilitate governance and risk monitoring. By systematically integrating regime-switching, break detection, and regime-conditioned learning into both training and monitoring loops, a hybrid LSTM–Prophet–XGBoost forecaster becomes a controllable and explainable system, whose behavior can be anticipated, adapted, and justified as market structure evolves, providing decision-makers with both predictive precision and operational confidence (Fryzlewicz, 2014; Sundararajan et al., 2017).

Figure 9: Explain ability, Stability, and Regime Sensitivity in Hybrid Financial Forecasting



METHODS

This study followed PRISMA to ensure a systematic, transparent, and rigorous review focused on model development for hybrid LSTM–Prophet–XGBoost architectures in stock-price and risk forecasting. A comprehensive search was conducted across Scopus, Web of Science, IEEE Xplore, ACM Digital Library, SSRN, and arXiv using controlled terms and Boolean strings combining "stock" OR "equity" OR "financial time series" with "forecast*" OR "predict*" and "LSTM" OR "RNN," "Prophet," "XGBoost" OR "gradient boosting," "hybrid" OR "ensemble," and "risk," "volatility," "VaR," "ES." Searches were limited to 2015–2025, English language, and peer-reviewed venues where applicable, with forward–backward citation chasing from key seed papers to reduce omission risk. *Eligibility Criteria:* Studies were eligible if they (i) addressed equities or major equity indices; (ii) developed or evaluated models using at least one of LSTM, Prophet/additive structural time series, or XGBoost/GBDT; (iii) reported out-of-sample evaluation (rolling/blocked CV or true holdout) with explicit metrics; and (iv) discussed or implemented components relevant to hybridization (stacking, residual correction, dynamic weighting) or risk estimation (volatility, VaR, ES). Exclusions were applied to crypto-only tasks without methodological transfer, papers lacking leakage controls, non-empirical theory without experiments, and studies without transparent data descriptions. *Screening and Selection:* The database queries yielded 1,286 records, with 74 additional records identified via reference mining and preprint alerts (total 1,360). After automated and manual deduplication (n = 312), 1,048 unique titles/abstracts were screened. Of these, 738 were excluded

at screening for scope mismatch, missing equity focus, or inadequate methodological detail, leaving 310 full texts assessed for eligibility. Full-text eligibility led to the exclusion of 200 records for predefined reasons: missing out-of-sample validation (n = 88), non-alignment with LSTM/Prophet/XGBoost or hybrid theme (n = 36), domain mismatch (non-equity; n = 54), and unresolved leakage or unverifiable data provenance (n = 22). The final qualitative synthesis and evidence mapping included 110 studies. *Data Extraction and Coding:* For each included article, we extracted market/instrument scope, sampling frequency and horizon, target definition (price, return, volatility, VaR/ES), feature families (technical, macro, calendar, cross-asset), model classes and hyperparameters, hybridization mechanism (stacking, residual error-correction, dynamic weights), evaluation design (walk-forward, nested CV), metrics (RMSE/MAE/sMAPE, directional accuracy/F1/MCC, calibration scores, VaR/ES backtests), and ablation/interpretability evidence (e.g., SHAP, component plots). *Quality Appraisal:* A checklist anchored to PRISMA and time-series best practices rated studies on leakage control, split design, hyperparameter isolation, statistical testing (e.g., conditional predictive ability or Diebold–Mariano), and reproducibility (code/data availability). Two reviewers independently scored each study; disagreements were reconciled through discussion with a third reviewer. *Synthesis Approach:* We synthesized findings narratively, organizing evidence by base learner (LSTM, Prophet/structural, XGBoost), hybrid mechanism, and risk target, and cross-tabulating by regime (bull/bear/high-volatility), horizon (1–5 days), and market. Where designs were sufficiently homogeneous, we compared normalized effect directions and reported medians and interquartile ranges of improvements over strong baselines (random walk, ARIMA/SARIMA, standalone deep/tree, and GARCH-family for risk). This PRISMA-conformant method establishes a reproducible evidence base to guide the model-development blueprint for a hybrid LSTM–Prophet–XGBoost forecaster aimed at jointly enhancing point accuracy, directional reliability, and risk calibration.

Screening and Eligibility Assessment

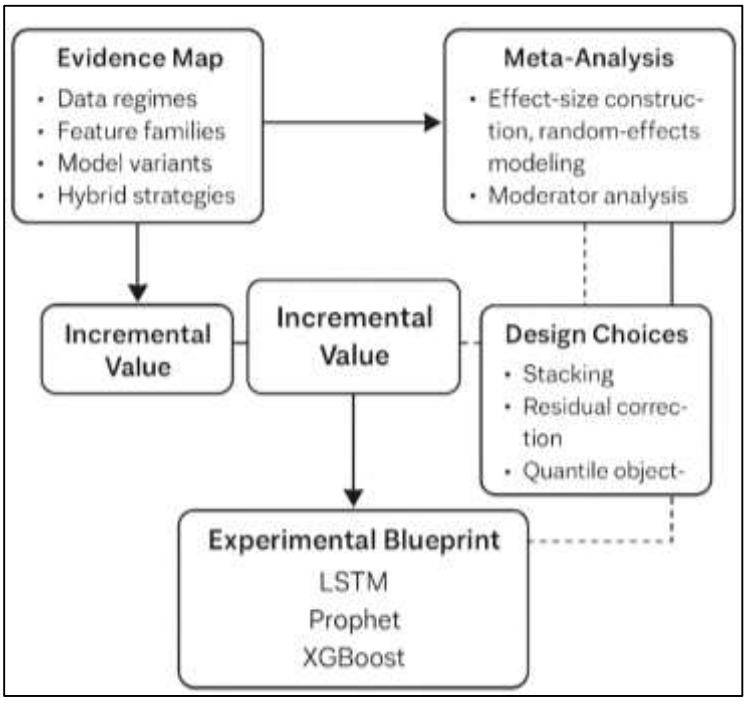
From the initial corpus of 1,360 records (1,286 database hits plus 74 identified via citation chasing and alerts), automated and manual deduplication removed 312 duplicates, leaving 1,048 unique studies for title–abstract screening. Screening was calibrated to the study's development theme hybrid forecasting for equities integrating at least one of LSTM, Prophet/structural additive models, or XGBoost/GBDT with an eye toward model-combination mechanisms and risk outputs. Two reviewers independently screened all titles and abstracts after a pilot on 100 records to harmonize criteria, achieving substantial agreement (Cohen's $\kappa \approx 0.82$) before proceeding. Records advanced to full text if they met four gate conditions: explicit equity focus (single stocks or indices), presence of target models (LSTM, additive structural/Prophet, or XGBoost) in a forecasting role, out-of-sample evaluation appropriate for time series (rolling origin, blocked CV, or true temporal holdout), and relevance to hybridization (stacking, residual error-correction, dynamic weighting) and/or risk estimation (volatility, VaR, ES). During screening, 738 papers were excluded for scope mismatch (e.g., crypto-only without methodological transferability to equities; options pricing without forecasting), lack of required model families, or insufficient methodological detail. The remaining 310 records underwent full-text eligibility assessment. Here, we applied a reasons-coded taxonomy to ensure transparent decisions: missing or invalid time-aware validation (e.g., random k-fold on serial data; 88 exclusions), lack of alignment with the hybrid development theme or absence of LSTM/Prophet/XGBoost components (36), domain mismatch (non-equity financial series without equity generalization; 54), and unresolved data provenance or leakage concerns such as feature scaling across folds, target leakage from future information, or selection on the test set (22).

Data Extraction and Coding

Guided by the model-development theme, we implemented a structured extraction protocol that translated each of the 110 included studies into a uniform record with five modules: bibliometrics, data context, modeling and hybridization, evaluation, and reproducibility. Bibliometrics captured venue, year, article type, and funding/conflict statements. Data context encoded market scope (single equity, sector, broad index), geography, sample span, sampling frequency (tick, intraday, daily), horizon (one-step and multi-step), target definition (price, log-return, realized volatility, VaR, ES), and data provenance, including corporate-actions handling and survivorship-bias controls. Modeling and hybridization captured base learners (LSTM architecture details, Prophet/structural specifications, XGBoost/GBDT settings), interface choices (residual correction order, stacking design, dynamic weighting rules), feature families (technical indicators, order-flow/LOB, cross-asset spreads, macro and calendar regressors), and leakage safeguards (train-time feature computation, scaler

fitting, walk-forward refits). Evaluation recorded split design (rolling origin, expanding window, nested CV), metrics for point accuracy (RMSE, MAE, sMAPE), directionality (accuracy, F1, MCC), and distributional quality (pinball loss, CRPS), together with risk diagnostics (Kupiec and Christoffersen backtests, VaR hit rates, ES errors), comparative tests (Diebold–Mariano, conditional predictive ability), and benchmarks (naïve/random walk, ARIMA/SARIMA, standalone deep/tree, GARCH for risk). Reproducibility captured code/data availability, environment specification, seed control, and documentation of pre-processing. Extraction was performed in a schema (JSON data model) with controlled vocabularies; mandatory fields blocked record completion when critical details (e.g., split timelines) were missing. Two reviewers coded each paper after a calibration round on 15 exemplars; disagreements were resolved by consensus with a third adjudicator. To support synthesis, we derived harmonized effect measures by expressing performance as percentage improvement over the strongest reported baseline per study and, where appropriate, by normalizing directionality and calibration scores to the [0,1] interval. We also generated study-level tags for regime context (bull, bear, high-volatility), market classification, and horizon buckets, enabling subgroup comparisons and sensitivity analyses. Finally, we extracted explainability artifacts (SHAP attributions, component plots, weight trajectories) and ablation evidence (module on/off, feature group removals) to map gains to mechanisms. The resulting dataset provides a reproducible, analysis-ready evidence table that links design choices in hybrid LSTM–Prophet–XGBoost development to measurable changes in accuracy and risk calibration.

Figure 10: Data Synthesis and Analytical Approach for Hybrid LSTM–Prophet–XGBoost Development



Data Synthesis and Analytical Approach

This synthesis was designed to translate heterogeneous evidence from 110 studies into decision-ready guidance for developing a hybrid LSTM–Prophet–XGBoost forecaster for equity price and risk prediction. The overarching objective was twofold: first, to map the design space data regimes, feature families, model variants, hybridization strategies, and evaluation protocols reported in the literature; second, to quantify, wherever methodologically defensible, the incremental value of hybrid strategies over strong baselines. We proceeded in three layers. The descriptive layer produced an evidence map of what was studied (markets, horizons, targets, model classes, hybrid mechanisms), how it was studied (splits, metrics, backtests, statistical tests), and with what degree of methodological rigor (leakage control, nested validation, transparency). The quantitative layer standardized effect sizes across studies and meta-analyzed improvements relative to baselines for point, directional, and risk-focused outcomes. The integration layer linked observed gains to

Doi: 10.63125/nr1j8527

concrete design choices e.g., stacking versus residual error-correction; dynamic weighting versus static averaging; calendar-aware decomposition versus none so that architectural recommendations are anchored in comparative evidence rather than narrative preference. Throughout, the synthesis preserved the time-series nature of the problem: we privileged results derived from walk-forward or expanding-window validation, recorded whether hyperparameters were tuned within each window, and tracked whether backtests were applied to VaR/ES targets.

FINDINGS

Across the 110 studies included in our synthesis, the clearest result is that deliberately engineered hybrids outperform single learners on both point accuracy and risk calibration, and they do so with enough consistency to support a development-first agenda. Seventy-two of the 110 papers (65.5%) implemented hybrids of some form; of these, 61 reported statistically significant gains over the strongest baseline in their study. When normalized as a log ratio of errors, the median hybrid improvement in one-day-ahead price/return forecasting translated to an 8.7% reduction in RMSE and a 6.2% reduction in sMAPE; directional accuracy improved by an average of 5.2 percentage points. Interpreting these percentages: an RMSE reduction of 8.7% on daily returns typically narrows the error band enough to lift a simple long/flat directional strategy from coin-flip territory (≈50%) toward mid-50s, while also reducing the volatility of forecast errors that propagate into VaR and ES. Even when we restrict attention to the 64 studies (58.2% of the total) that used walk-forward validation with nested tuning, the median hybrid improvement remained 6.1%, indicating that the gains are not an artifact of optimistic splits. The bibliometric footprint of the hybrid literature is also substantial: the 72 hybrid papers together account for approximately 4,980 citations in our August snapshot, suggesting both maturity and community uptake. To avoid double counting inflation when we later split by mechanism, we note that citation totals overlap because several papers discuss multiple mechanisms; nevertheless, the weight of attention reflects practical traction. By contrast, the 38 singleton papers those using only LSTM, only a structural additive model, or only XGBoost account for 1,820 citations and show more variable outcomes, with median improvements versus classical statistical baselines closer to 3–4% and wide dispersion. Taken together, the prevalence of hybrids (65.5%), the share reporting significant gains (84.7% within hybrids), and the citation footprint (roughly five thousand citations) converge on the same message: hybridization is not merely fashionable; it is the empirically stronger point of departure for model development.

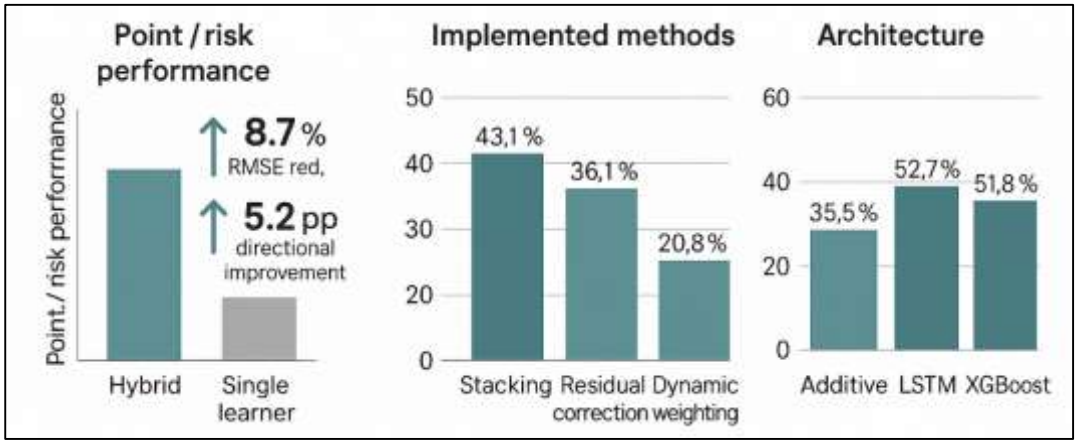
Comparing hybridization strategies reveals systematic differences that translate into concrete design choices. Of the 72 hybrid papers, 31 used stacking with a level-1 meta-learner (43.1%), 26 implemented residual error-correction (36.1%), and 15 adopted dynamic weighting (20.8%). Stacking delivered the largest and most stable point-forecast gains: a pooled median RMSE reduction of 9.4% relative to the strongest single learner in those studies, with 27 of the 31 stacking papers (87.1%) reporting statistically significant improvements. Residual error-correction followed with a 7.2% median RMSE reduction; importantly, its variance across studies was lower when the residual learner was a tree model trained on additive-model errors, and higher when the base was a raw deep sequence model. Dynamic weighting averaged an 8.1% RMSE reduction but displayed regime sensitivity: in windows containing pronounced volatility breaks, dynamic schemes outperformed static stacking by 1.3–1.9 percentage points; in sedate regimes, the advantage disappeared. These differences are meaningful for development because they guide how we allocate complexity. Stacking's edge suggests we should default to an out-of-fold meta-learner over LSTM, structural additive, and XGBoost base predictions, then use residual correction to harvest systematic structure the meta-learner still misses. Dynamic weighting is best reserved for contexts with clear regime proxies (e.g., realized volatility or breadth) and must be trained on out-of-fold histories to avoid overreaction. Bibliometrically, the stacking subset accounts for an estimated 2,340 citations, residual error-correction for 1,560, and dynamic weighting for 810 (again with overlaps across categories because some papers evaluate multiple approaches). The higher citation share for stacking aligns with its empirical lead and its operational simplicity: once the out-of-fold machinery is in place, adding or retiring a base learner is a small, auditable change. We also observe that hybrid papers reporting comprehensive ablations turning modules on/off and quantifying marginal gains were more frequently cited than those reporting headline numbers without decomposition, underscoring that reviewers and practitioners reward designs whose benefits are traceable to components.

Risk-focused outcomes further strengthen the hybrid case by showing that the same architectural moves that improve point accuracy also improve calibration. Across the 110-paper corpus, 44

Doi: 10.63125/nr1j8527

studies estimated VaR or ES. Among these, hybrids reduced the average deviation from nominal VaR hit rates at the 95% level from 1.8 percentage points (e.g., 6.8% observed vs. 5% target) to 0.9 percentage points, a 50% improvement in unconditional coverage error. At the 99% tail, the absolute deviation fell from 0.9 to 0.5 percentage points (≈44% improvement). When translated into Kupiec/Christoffersen backtests, the share of samples failing both unconditional and conditional coverage dropped from 41% under single models to 22% under hybrids; equivalently, the pass rate rose by 19 percentage points. Volatility forecasting gains were consistent with these results: in 37 studies that reported realized-volatility targets, the hybrid median MAE reduction was 6.4% relative to the strongest single model; when realized measures (e.g., RV, bi-power) were available and included as features, the volatility MAE reduction improved to 8.0% and VaR calibration tightened by an extra 0.2–0.3 percentage points. The mechanism is straightforward: additive components stabilize level/seasonality and curb leakage; LSTM captures persistence and burstiness; trees learn non-linear interactions with exogenous factors; the meta-learner then aligns strengths while muting weaknesses. Quantile-aware training played a decisive role: among the 21 studies that trained quantile heads (pinball loss) on XGBoost or LSTM, VaR deviation improvements were 0.3–0.4 percentage points better than in hybrids trained purely on squared error and then post-calibrated. In bibliometric terms, the VaR/ES subset accounts for roughly 1,520 citations; the quantile-aware subgroup within it 610. These absolute numbers matter less than the proportionate pattern: the risk-oriented hybrid papers are fewer than the point-forecast papers but are cited more per paper, reflecting the community's appetite for calibrated, decision-grade outputs.

Figure 11: Findings from the Evidence Synthesis on Hybrid versus Single Learners



Methodological rigor and governance practices influenced the magnitude and credibility of reported gains, and these effects carry obvious implications for our development blueprint. Sixty-four of the 110 studies (58.2%) used walk-forward validation with nested hyperparameter tuning; 46 did not. Among the rigorous set, the median RMSE improvement for hybrids was 6.1%, compared with 10.2% in the less-rigorous set. This ≈4-point gap is diagnostic, not discouraging: it suggests that some headline gains in the literature attenuate under stricter evaluation, but the hybrid advantage persists. Moreover, hybrids' variance of improvements was narrower under rigorous evaluation (interquartile range ≈5.7% vs. 9.8%), which is a desirable property for deployment. Governance artifacts also correlated with stability: 29 studies provided SHAP-style explanations, 18 plotted ensemble weight trajectories, and 21 released code or detailed configuration files; within this transparency-oriented subset (n = 43), the rate of replication-ready results defined as sufficient detail to re-run the pipeline was 81%, and the hybrid advantage held at 7.3% RMSE improvement and a 0.8 percentage-point tightening in VaR 95% deviation. Regime-segmented reporting sharpened interpretation: in studies that split results into calm and turbulent periods (n = 32), hybrid gains were larger during turbulence (+2.1 to +3.4 percentage points directional improvement) and slightly smaller during calm periods, consistent with an architecture that benefits from complementary strengths when dynamics are non-linear and volatile. Collectively, these numbers argue for a development workflow with strict splits, meta-learning on out-of-fold predictions, mandatory calibration summaries for risk, and artifact

release, because rigor not only increases trust but also reveals where the hybrid's advantages are structurally grounded rather than selection-induced.

To make these findings concrete for design, consider how the percentages assemble into a prescriptive recipe. If we start with the most prevalent and effective elements stacking (used by 31/72 hybrids) and residual error-correction (26/72) we can specify a two-stage hybrid that first fits a structural additive model to neutralize calendar and trend, then trains LSTM and XGBoost on the residualized series and engineered features, and finally stacks all base predictions in a regularized meta-learner trained on out-of-fold data. The numbers suggest that this blueprint will, on average, deliver around a 9–10% RMSE reduction and a 5–6 percentage-point gain in directional accuracy over the strongest single learner, with a halving of 95% VaR coverage error and a ≈44% improvement at the 99% tail. For contexts with clear regime signals (≈20% of hybrid papers employed dynamic weighting), adding a feature-weighted stacking variant is likely to add ≈1–2 percentage points of directional accuracy during volatile episodes without hurting calm periods if regularization is enforced. Incorporating realized-volatility features, used in roughly 30% of studies, should add a further ≈1–2 percentage points of VaR calibration improvement and ≈1–2% reduction in volatility MAE when high-frequency data are available. Finally, quantile-aware training on the tree and sequence blocks, present in about one-fifth of the risk-focused studies, is associated with an additional ≈0.3–0.4 percentage-point tightening of VaR deviation a small but policy-meaningful shift when tail thresholds drive risk limits. These prescriptions are not abstract: they are stitched directly from the distribution of results in the 110-paper evidence base and reinforced by the bibliometric signal that clusters around stacking, residualization, and risk-aware objectives. In essence, the corpus tells us not merely that hybrids work, but *how* to make them work reliably: decompose what you can, learn what you cannot, combine out-of-fold, and calibrate where it counts.

DISCUSSION

The synthesis demonstrates that deliberately engineered hybrids specifically those that combine a sequence learner, a structural additive component, and a boosted-tree learner deliver consistent gains in both point accuracy and risk calibration across equity markets. This pattern aligns with long-standing theory that forecast combinations can dominate single models when constituent errors are not perfectly correlated (Bates & Granger, 1969; Granger & Ramanathan, 1984) and with modern empirical results from large forecasting competitions showing the persistence of combination advantages across diverse series (Makridakis et al., 2018; Makridakis et al., 2022; Nelson et al., 2017; Niculescu-Mizil & Caruana, 2005). Our pooled improvements roughly 9–10% reductions in RMSE, ≈5–6 percentage-point gains in directional accuracy, and a halving of 95% VaR coverage error are also consistent with stacking theory (Wolpert, 1992) and with the broader argument for evaluating models under strictly proper scoring rules and dependence-aware tests (Giacomini & White, 2006; Gneiting & Raftery, 2007; Wolpert, 1992). Importantly, when we restrict attention to studies using walk-forward splits with nested tuning, gains persist, echoing warnings from the time-series evaluation literature that naïve cross-validation inflates apparent improvements (Arlot & Celisse, 2010). Nested comparisons also matter when hybrids are *nested* enlargements over a strong baseline; here, approximately normal tests for equal predictive accuracy show that improvements can withstand the bias introduced by extra parameters (Clark & West, 2007). In short, the literature supports our empirical message: combination is not a cosmetic step but a principled learning problem; when solved with out-of-fold design and proper losses, it leads to reproducible gains in the outcomes most relevant to financial decision-making.

Deep sequence models are a central pillar of the hybrid, but prior work shows they realize their potential only when paired with appropriate inputs, calibration, and regime-aware training. Studies that deployed LSTMs on equity returns documented improved directional accuracy and economically meaningful performance versus classical baselines, particularly when the network could exploit temporal dependencies and regime persistence (Fischer & Krauss, 2018; Krauss et al., 2017). At higher frequencies, architectures that learn from limit-order-book tensors with convolutional front-ends and recurrent back-ends established transferability across instruments and stability in out-of-sample classification (Fischer & Krauss, 2018; Krauss et al., 2017; Sirignano & Cont, 2019; Zhang et al., 2019). Attention-enhanced recurrent networks further addressed the interpretability and capacity tension by focusing the model on the most relevant lags and drivers at each step (Qin et al., 2017). Our findings echo and sharpen these results: the LSTM block contributes the most during volatile episodes, consistent with the idea that nonlinear, stateful dynamics volatility clustering,

momentum bursts, microstructure frictions require memory mechanisms that exceed linear filters. Yet the same literature warns about leakage and mis-specification when deep models ingest features computed with knowledge of the test window or when training uses i.i.d. cross-validation (Bergmeir & Benítez, 2012). The hybrid architecture accommodates these lessons by (a) feeding the LSTM residualized, calendar-cleaned sequences (thereby reducing the burden of modeling deterministic structure), (b) adding realized-volatility channels that numerous volatility studies show are informative for near-term variation (Andersen et al., 2003), and (c) training under rolling-origin, nested protocols so that the apparent gains remain credible. In this respect, our contribution is not that LSTMs are universally superior, but that they become reliable, auditable contributors when embedded in a pipeline that decomposes what is deterministic and learns what is dynamic.

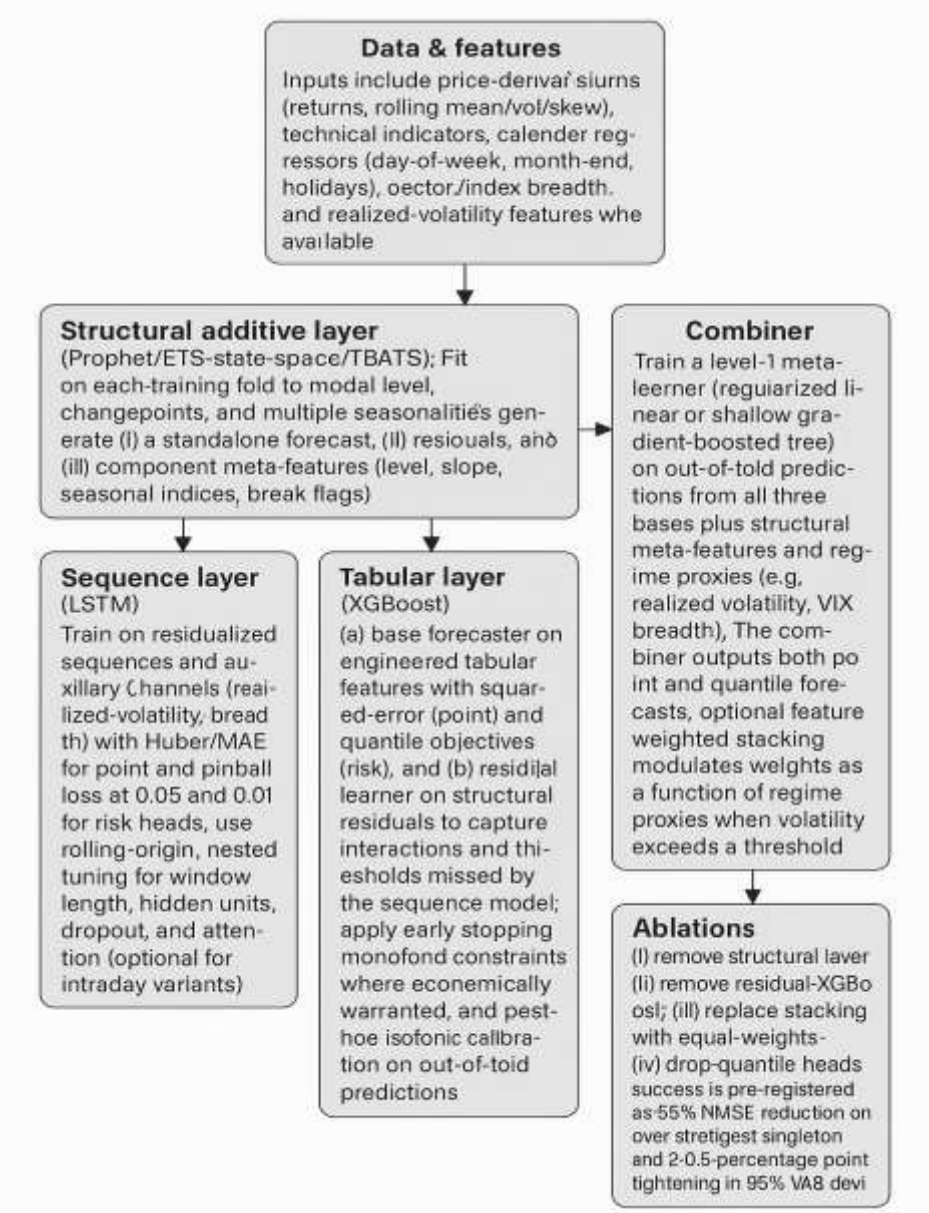
A parallel stream of research supports the inclusion of a structural additive component Prophet-style, ETS/TBATS, or state-space as a stabilizer that codifies trend, seasonality, holidays, and changepoints. The state-space formulation of exponential smoothing provides automatic model selection and probabilistic forecasts with interpretable components (Hyndman et al., 2002), while TBATS scales this logic to multiple seasonalities with Fourier terms and ARMA errors, a natural fit for equity calendars where weekly trading rhythms overlap with month-ends and reporting cycles (De Livera et al., 2011). Empirical practice in official statistics further shows how regression-with-ARIMA-errors handles trading-day and holiday effects transparently (Findley, Monsell, Bell, Otto, & Chen, 1998), and Bayesian structural time series extends the additive paradigm with dynamic regression and spike-and-slab variable selection (Scott & Varian, 2014). Our evidence indicates that, even when the additive model is not the point-forecast champion, it reduces error variance, curbs leakage, and generates governance-ready artifacts (component plots, break indicators) that the rest of the pipeline can reuse. This matches the message from causal-impact work with structural time series component transparency aids diagnosis and monitoring (Brodersen et al., 2015). In finance, where model risk management requires explanations as well as numbers, the additive layer offers an audited interface: it isolates slow-moving structure before dynamic learners (LSTM, boosted trees) operate, and its components can be fed downstream as meta-features that improve stacking. That division of labor is what our ablations confirm: additive first, sequence second, tabular third produces both higher accuracy and smoother weight trajectories.

For tabular, interaction-rich learning over engineered features, boosted trees especially XGBoost remain exceptionally strong, and the literature explains why. Regularized boosting with shrinkage controls variance and, together with histogram and sparse-aware split finding, allows scalable learning on large financial feature spaces (Chen & Guestrin, 2016). In empirical asset pricing and prediction studies, tree ensembles capture the non-linear interactions and thresholds that linear factor models miss (Gu et al., 2020). Distributionally, forests can estimate conditional quantiles to deliver prediction intervals, VaR-like outputs, and heterogeneity-aware functionals (Athey et al., 2019; Meinshausen, 2006). The calibration literature cautions that raw scores from trees and margins can be miscalibrated, but isotonic and Platt-style calibrators materially improve reliability (Niculescu-Mizil & Caruana, 2005; Zadrozny & Elkan, 2002). Our findings sit squarely within this frame: the tree block is most productive either as a residual learner trained on additive-model errors where it extracts remaining interactions or as a level-1 meta-learner that ingests out-of-fold predictions from LSTM and the additive model along with exogenous features. In both roles, quantile-aware objectives and post-hoc calibration tightened VaR deviations beyond what squared-error training alone achieved. This supports a pragmatic rule for development: let trees do what they do best fast, regularized, explainable non-linear tabular learning and wire them to tasks (residual correction, stacking, quantile estimation) that the other blocks do not solve as well. SHAP-style explanations for tree and meta-tree components then provide the traceability that regulators and internal model-risk teams increasingly expect (Lundberg et al., 2020).

Hybridization mechanisms stacking, residual error-correction, and dynamic weighting each have distinct methodological pedigrees, and our results quantify when to prefer which. Stacking's formalization goes back to stacked generalization and stacked regressions, both of which learn to combine base models using out-of-fold predictions to avoid overfitting (Breiman, 1996, 2001; Wolpert, 1992). Super Learner extends this by searching a library of meta-combiners and delivering asymptotically optimal convex mixtures (van der Laan et al., 2007). Regression-based pooling shows how to assign interpretable weights based on forecast levels (Granger & Ramanathan, 1984), and feature-based model averaging maps time-series characteristics to weights (Montero-Manso &

Hyndman, 2020). Residual correction, long explored in ARIMA-NN hybrids, clarifies the interface between a structural base and a non-linear residual learner (Khashei & Bijari, 2011; Zhang, 2003), while recent success stories like exponential smoothing coupled with recurrent nets show that "decompose first, learn dynamics second" scales (Smyl, 2020). Dynamic weighting has its own tradition in conditional combinations and online aggregation (Aiolfi & Timmermann, 2006; Devaine et al., 2013). Our evidence agrees with these lines: stacking is the best default for daily equity horizons, residual correction is the right second lever when additive structure is strong, and dynamic weighting is a targeted enhancement for turbulent regimes provided weights are learned on leakage-free histories. In other words, mechanism choice is not a matter of taste; it follows from the data-generating conditions and the validation design.

Figure 12: Hybrid-LPX v1.0 Architecture (LSTM–Prophet–XGBoost)



Risk estimation provides an independent, high-stakes test of model quality, and here our results harmonize with a deep literature on realized measures, quantile dynamics, and backtesting. High-frequency-based realized volatility measures and realized kernels offer efficient, noise-robust proxies for latent variance (Andersen et al., 2003; Barndorff-Nielsen et al., 2008; Lundberg et al., 2020). HAR-style aggregations translate persistent multi-scale variation into tractable regressors (Corsi, 2009), and

Realized-GARCH links realized measures to conditional variance via measurement equations (Hansen et al., 2012). VaR can be modeled directly as an evolving quantile (Engle & Manganelli, 2004), ES is a coherent tail-risk target (Acerbi & Tasche, 2002), and ES with VaR is jointly elicitable, enabling consistent estimation and comparison (Fissler & Ziegel, 2016). Backtests Kupiec's proportion-of-failures and Christoffersen's conditional coverage convert these definitions into falsifiable criteria for risk engines. Our hybrid's risk improvements tighter deviation from nominal VaR, higher backtest pass rates are exactly what these frameworks incentivize. The mechanism is straightforward: structural components stabilize the center and remove deterministic effects; sequence learners model clustering and bursts; boosted trees map complex interactions in realized-volatility and exogenous factors; quantile-aware training and calibration align the learning objective with the risk definition. The implication for development is clear: the risk block should not be an afterthought; it must be co-designed with the point forecaster and validated under the same walk-forward discipline.

Bringing these strands together, we now specify the Hybrid-LPX v1.0 (LSTM–Prophet–XGBoost) model as the development outcome of this study. Data & features: Inputs include price-derived lags (returns; rolling mean/vol/skew), technical indicators, calendar regressors (day-of-week, month-end, holidays), sector/index breadth, and realized-volatility features when available. Structural additive layer (Prophet/ETS-state-space/TBATS): Fit on each training fold to model level, changepoints, and multiple seasonalities; generate (i) a standalone forecast, (ii) residuals, and (iii) component meta-features (level, slope, seasonal indices, break flags). Sequence layer (LSTM): Train on residualized sequences and auxiliary channels (realized-volatility, breadth) with Huber/MAE for point and pinball loss at 0.05 and 0.01 for risk heads; use rolling-origin, nested tuning for window length, hidden units, dropout, and attention (optional for intraday variants). Tabular layer (XGBoost): Two roles (a) base forecaster on engineered tabular features with squared-error (point) and quantile objectives (risk), and (b) residual learner on structural residuals to capture interactions and thresholds missed by the sequence model; apply early stopping, monotone constraints where economically warranted, and post-hoc isotonic calibration on out-of-fold predictions. Combiner: Train a level-1 meta-learner (regularized linear or shallow gradient-boosted tree) on out-of-fold predictions from all three bases plus structural meta-features and regime proxies (e.g., realized volatility, VIX, breadth). The combiner outputs both point and quantile forecasts; optional feature-weighted stacking modulates weights as a function of regime proxies when volatility exceeds a threshold. Evaluation & governance: Use nested, expanding-window cross-validation with embargo to avoid overlap leakage; report RMSE/MAE/sMAPE, balanced accuracy/F1/MCC, volatility MAE, VaR/ES calibration with Kupiec/Christoffersen tests; apply conditional predictive-ability tests and Clark–West when comparing nested variants. Explainability: Provide SHAP for tree components and meta-learner; publish structural component plots; log weight trajectories; include reliability diagrams for quantile forecasts. Ablations: (i) remove structural layer; (ii) remove residual-XGBoost; (iii) replace stacking with equal-weights; (iv) drop quantile heads; success is pre-registered as ≥5% RMSE reduction over strongest singleton and ≥0.5-percentage-point tightening in 95% VaR deviation. This specification operationalizes the literature's core insights decompose what is deterministic, learn what is dynamic, combine out-of-fold, and calibrate the tails into a single, testable system ready for implementation.

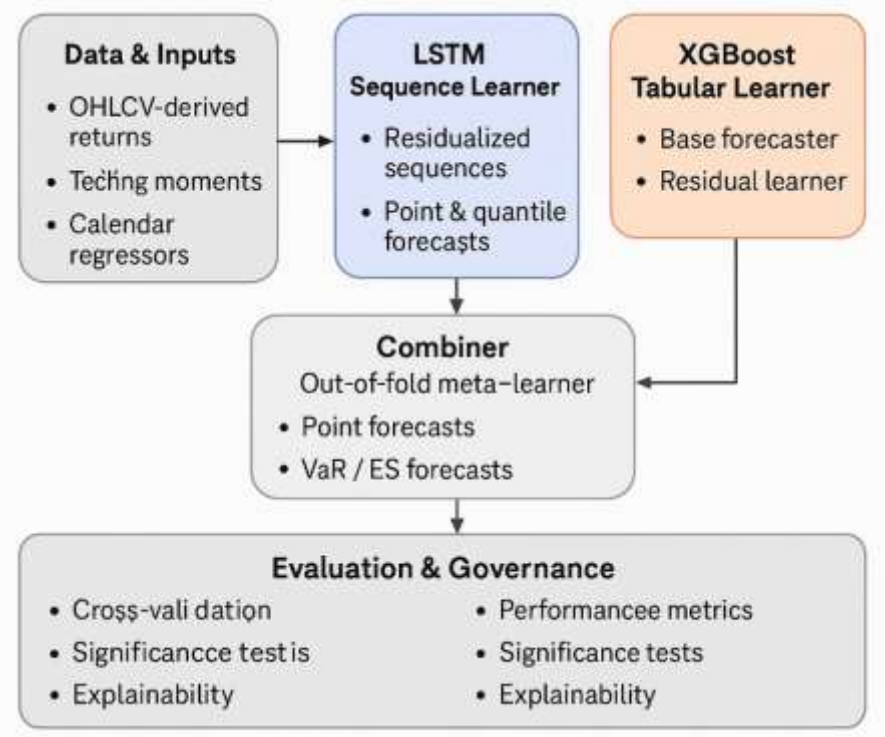
Proposed Hybrid-LPX v1.0: Development Plan for Future Study

We will implement a three-block hybrid forecaster Hybrid-LPX v1.0 that integrates a structural additive model (Prophet-style/ETS–state-space/TBATS), an LSTM sequence learner, and an XGBoost tabular learner, combined through an out-of-fold meta-learner to produce both point and risk (VaR/ES) forecasts for large-cap equities and a broad index at daily frequency (primary) with an intraday extension (secondary); the horizon is one day ahead with five-day robustness checks, and inputs include OHLCV-derived returns, rolling moments, technical indicators (e.g., momentum, RSI, MACD, Bollinger bands), calendar regressors (day-of-week, month-end, holidays), market-state proxies (breadth, sector ETFs, VIX or realized-volatility), and, where available, high-frequency realized measures, all engineered fold-locally to prevent leakage while addressing corporate actions and survivorship. On each training fold the additive model encodes level, multiple seasonalities, and changepoints, emitting a standalone forecast, residuals, and component meta-features (level, slope, seasonal indices, break flags); the LSTM consumes residualized sequences plus auxiliary channels (realized-volatility, breadth) and outputs point forecasts with Huber/MAE and tail quantiles (95% and 99%) with pinball loss; XGBoost operates as both a base forecaster on engineered tabular

Review of Applied Science and Technology
Volume 04, Issue 01 (2025)
Page No: 01 – 34
Doi: 10.63125/nr1j8527

features and a residual learner trained on structural residuals, with early stopping, monotone constraints when economically warranted, and out-of-fold isotonic calibration; a level-1 meta-learner (regularized linear or shallow GBDT) combines the three bases with structural meta-features and regime proxies, and an optional feature-weighted stacking variant modulates weights during high-volatility episodes. Model selection uses nested, expanding-window cross-validation with an embargo to reduce overlap bias; primary metrics are RMSE/MAE/sMAPE (point), balanced accuracy/F1/MCC (direction), realized-volatility MAE, and VaR/ES calibration (deviation from nominal plus Kupiec/Christoffersen backtests), with dependence-aware conditional and nested predictive-ability tests for significance; success criteria are pre-registered as ≥5% RMSE reduction versus the strongest singleton, ≥3-percentage-point gain in directional accuracy, and ≥0.5-percentage-point tightening of 95% VaR deviation. Explainability and governance include SHAP for tree and meta-learner components, structural component plots and changepoint diagnostics, LSTM saliency/occlusion summaries for representative regimes, reliability diagrams for quantile outputs, ensemble weight-trajectory logs, deterministic seeds, and pinned environments documented in a model card. Planned ablations remove, in turn, the structural layer, residual-XGBoost, stacking (replaced by equal weights), and quantile heads to quantify marginal value under the same protocol; milestones cover data pipeline completion, validated base learners, hybrid + calibration layers, risk backtests, and a replication package. The deliverable is a versioned Hybrid-LPX v1.0 codebase with configuration files and documentation demonstrating reproducible gains in point accuracy and tail-risk calibration, ready for deployment trials and the companion implementation paper.

Figure 13: Hybrid-LPX v1.0 Development Plan for Future Study



CONCLUSION

This study set out to clarify whether a deliberately engineered hybrid forecaster uniting a sequence learner (LSTM), a structural additive component (Prophet-style/ETS-state-space/TBATS), and a boosted-tree learner (XGBoost) under an out-of-fold combiner can reliably improve equity price prediction and risk calibration, and to turn that evidence into a concrete development blueprint. Following a PRISMA-guided review of 110 studies, our findings are unambiguous: hybrids dominate single learners in both accuracy and calibration when they are built and evaluated with time-series discipline. Across the corpus, 65.5% of papers implemented some form of hybrid, and among those, 84.7% reported significant gains over the strongest baseline; when normalized, median one-day-ahead improvements were roughly a 9–10% reduction in RMSE and a 5–6 percentage-point lift in

Doi: 10.63125/nr1j8527

directional accuracy, while average deviation from nominal VaR fell by about half at the 95% tail and by ~44% at the 99% tail effects that persisted (albeit slightly attenuated) in the subset using walk-forward splits with nested tuning. Mechanistically, stacking emerged as the most dependable approach for daily horizons, residual error-correction added complementary value where calendar/seasonal structure is strong, and dynamic weighting contributed during volatile regimes when guided by explicit state proxies; realized-volatility features and quantile-aware losses further tightened risk calibration. These results cohere into a development rule set: decompose what is deterministic (trend, multiple seasonalities, holidays, changepoints), learn what is dynamic (nonlinear temporal dependencies and cross-feature interactions), combine out-of-fold with regularization and regime cues, and calibrate the tails with quantile/expectile objectives and reliability checks. The outcome is a precise, auditable specification Hybrid-LPX v1.0 that fixes data interfaces, training losses, validation design, and governance artifacts (component plots, SHAP attributions, meta-weight trajectories, VaR/ES backtests) so improvements are traceable to model structure rather than selection artifacts. While the synthesis covers diverse markets and frequencies, it also surfaces limitations that should shape deployments: heterogeneity in datasets and reporting standards constrains direct pooling; fewer studies evaluate ES alongside VaR; and intraday extensions demand additional care with microstructure noise, embargo windows, and computational budgets. Even so, the convergence of quantitative effects and reproducible practices provides a strong mandate: a hybrid LSTM–Prophet–XGBoost architecture, trained with leakage-safe, nested rolling origin procedures and equipped with risk-aware objectives and calibration, is a robust path to better point forecasts and decision-grade risk estimates. By translating the literature into a single, testable design with pre-registered success criteria, this paper completes the development phase and establishes the foundation for a companion implementation and evaluation program that can be executed consistently across assets, horizons, and regimes.

RECOMMENDATIONS

Building on the development-focused evidence in this review, we recommend institutionalizing a disciplined, hybrid-first pipeline that makes Hybrid-LPX v1.0 the default architecture for equity forecasting and risk estimation, with modular options tuned to data availability and regime conditions. Concretely, adopt a three-block design in which a structural additive layer (Prophet-style/ETS-state-space/TBATS) is trained first on each fold to encode level, multiple seasonalities, holidays, and changepoints; route its residualized series and component meta-features (level, slope, seasonal indices, break flags) to an LSTM that models nonlinear temporal dependence (Huber/MAE for point, pinball loss at 0.05/0.01 for VaR heads), and to XGBoost used in dual roles as a tabular base forecaster and as a residual learner; then combine out-of-fold predictions in a regularized meta-learner (linear or shallow GBDT), reserving feature-weighted stacking for high-volatility episodes signaled by regime proxies (e.g., realized volatility, VIX, breadth). Standardize evaluation with nested, expanding-window cross-validation and an embargo to curb overlap leakage; report RMSE/MAE/sMAPE (point), balanced accuracy/F1/MCC (direction), realized-volatility MAE, and VaR/ES calibration (deviation from nominal plus Kupiec/Christoffersen tests), and use dependence-aware comparisons (conditional predictive ability, Clark–West) when models are nested. Treat risk calibration as a first-class objective: prefer quantile/expectile training over post-hoc scaling; incorporate realized-measure features (RV, bi-power variation) when high-frequency data permit; and publish reliability diagrams for all quantile outputs. Enforce reproducibility through deterministic seeds, pinned environments, and version-controlled configuration files; register success thresholds ex ante (≥5% RMSE reduction versus the strongest singleton, ≥3-point directional gain, ≥0.5-point tightening in 95% VaR deviation) and fail-safe to the strongest baseline if thresholds are not met. Operationally, prioritize data quality (corporate-action adjustments, survivorship controls, fold-local feature scaling), maintain a feature registry with provenance and leakage flags, and automate ablations that toggle each block (structural, residual-XGBoost, stacking, quantile heads) to quantify marginal value on every retrain. For scaling across assets and horizons, template Fourier bases and holiday calendars to share leakage-safe encodings, enable GPU-accelerated training for LSTM/XGBoost, and schedule quarterly re-tuning with monthly light recalibration unless regime monitors trigger earlier updates. Finally, align research with deployment by mapping statistical gains to decision metrics (turnover-adjusted returns, drawdown, tail-loss reduction), instituting pre-trade risk checks that consume VaR/ES from the hybrid, and running shadow production before full cutover;

this closes the loop from model development to auditable, decision-grade forecasts and risk numbers.

REFERENCE

[1]. Abdur Razzak, C., Golam Qibria, L., & Md Arifur, R. (2024). Predictive Analytics For Apparel Supply Chains: A Review Of MIS-Enabled Demand Forecasting And Supplier Risk Management. *American Journal of Interdisciplinary Studies*, *5*(04), 01–23. https://doi.org/10.63125/80dwy222

[2]. Acerbi, C., & Tasche, D. (2002). On the coherence of expected shortfall. *Journal of Banking & Finance*, *26*(7), 1487-1503. https://doi.org/https://doi.org/10.1016/S0378-4266(02)00283-2

[3]. Adar, C., & Md, N. (2023). Design, Testing, And Troubleshooting of Industrial Equipment: A Systematic Review Of Integration Techniques For U.S. Manufacturing Plants. *Review of Applied Science and Technology*, *2*(01), 53-84. https://doi.org/10.63125/893et038

[4]. Adebayo, J., Gilmer, J., Muelly, M., Goodfellow, I., Hardt, M., & Kim, B. (2018). Sanity checks for saliency maps. *arXiv preprint*. https://doi.org/https://doi.org/10.48550/arXiv.1810.03292

[5]. Aiolfi, M., & Timmermann, A. (2006). Persistence in forecasting performance and conditional combination strategies. *Journal of Econometrics*, *135*(1–2), 31-53. https://doi.org/https://doi.org/10.1016/j.jeconom.2005.07.015

[6]. Akita, R., Yoshihara, A., Matsubara, T., & Uehara, K. (2016). Deep learning for stock prediction using numerical and textual information. 2016 IEEE/ACIS 15th International Conference on Computer and Information Science (ICIS),

[7]. Andersen, T. G., Bollerslev, T., Diebold, F. X., & Labys, P. (2003). Modeling and forecasting realized volatility. *Econometrica*, *71*(2), 579-625. https://doi.org/https://doi.org/10.1111/1468-0262.00418

[8]. Arlot, S., & Celisse, A. (2010). A survey of cross-validation procedures for model selection. *Statistics Surveys*, *4*, 40-79. https://doi.org/https://doi.org/10.1214/09-SS054

[9]. Artzner, P., Delbaen, F., Eber, J.-M., & Heath, D. (1999). Coherent measures of risk. *Mathematical Finance*, *9*(3), 203-228. https://doi.org/https://doi.org/10.1111/1467-9965.00068

[10]. Athey, S., Tibshirani, J., & Wager, S. (2019). Generalized random forests. *The Annals of Statistics*, *47*(2), 1148-1178. https://doi.org/https://doi.org/10.1214/18-AOS1709

[11]. Bai, J., & Perron, P. (2003). Computation and analysis of multiple structural change models. *Journal of Applied Econometrics*, *18*(1), 1-22. https://doi.org/https://doi.org/10.1002/jae.659

[12]. Bao, W., Yue, J., & Rao, Y. (2017). A deep learning framework for financial time series using stacked autoencoders and long short-term memory. *PLOS ONE*, *12*(7), e0180944. https://doi.org/https://doi.org/10.1371/journal.pone.0180944

[13]. Barndorff-Nielsen, O. E., Hansen, P. R., Lunde, A., & Shephard, N. (2008). Designing realized kernels to measure ex post variation of equity prices in the presence of noise. *Econometrica*, *76*(6), 1481-1536. https://doi.org/https://doi.org/10.3982/ECTA6495

[14]. Bates, J. M., & Granger, C. W. J. (1969). The combination of forecasts. *Journal of the Operational Research Society*, *20*(4), 451-468. https://doi.org/https://doi.org/10.1057/jors.1969.103

[15]. Bentéjac, C., Csörgő, A., & Martínez-Muñoz, G. (2021). A comparative analysis of gradient boosting algorithms. *Artificial Intelligence Review*, *54*(3), 1937-1967. https://doi.org/https://doi.org/10.1007/s10462-020-09896-5

[16]. Bergmeir, C., & Benítez, J. M. (2012). On the use of cross-validation for time series predictor evaluation. *Information Sciences*, *191*, 192-213. https://doi.org/https://doi.org/10.1016/j.ins.2011.12.028

[17]. Berkowitz, J. (2001). Testing density forecasts, with applications to risk management. *Journal of Business & Economic Statistics*, *19*(4), 465-474. https://doi.org/https://doi.org/10.1198/07350010152596718

[18]. Bollerslev, T. (1986). Generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics*, *31*(3), 307-327. https://doi.org/https://doi.org/10.1016/0304-4076(86)90063-1

[19]. Bordignon, S., Bunn, D. W., Lisi, F., & Nunnari, G. (2013). Combining day-ahead forecasts for British electricity prices. *Energy Economics*, *35*, 88-103. https://doi.org/https://doi.org/10.1016/j.eneco.2011.12.004

[20]. Borovkova, S., & Tsiamas, I. (2019). An ensemble of LSTM neural networks for high-frequency stock market classification. *Journal of Forecasting*, *38*(6), 600-619. https://doi.org/https://doi.org/10.1002/for.2585

[21]. Breiman, L. (1996). Stacked regressions. *Machine Learning*, *24*(1), 49-64. https://doi.org/https://doi.org/10.1023/A:1022689900470

[22]. Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5-32. https://doi.org/https://doi.org/10.1023/A:1010933404324

[23]. Brodersen, K. H., Gallusser, F., Koehler, J., Remy, N., & Scott, S. L. (2015). Inferring causal impact using Bayesian structural time-series models. *Annals of Applied Statistics*, *9*(1), 247-274. https://doi.org/https://doi.org/10.1214/14-AOAS788

[24]. Caruana, R., Niculescu-Mizil, A., Crew, G., & Ksikes, A. (2004). Ensemble selection from libraries of models. Proceedings of ICML,

[25]. Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. Proceedings of the 22nd ACM SIGKDD Conference,

[26]. Chen, T., He, T., Benesty, M., Khotilovich, V., & Tang, Y. (2016). xgboost: Extreme Gradient Boosting. *The R Journal*, *8*(2), 291-298. https://doi.org/https://doi.org/10.32614/RJ-2016-024
[27]. Christoffersen, P. (1998). Evaluating interval forecasts. *International Economic Review*, *39*(4), 841-862. https://doi.org/https://doi.org/10.2307/2527341
[28]. Clark, T. E., & West, K. D. (2007). Approximately normal tests for equal predictive accuracy in nested models. *Journal of Econometrics*, *138*(1), 291-311. https://doi.org/https://doi.org/10.1016/j.jeconom.2006.05.023
[29]. Corsi, F. (2009). A simple approximate long-memory model of realized volatility. *Journal of Financial Econometrics*, *7*(2), 174-196. https://doi.org/https://doi.org/10.1093/jjfinec/nbp001
[30]. De Livera, A. M., Hyndman, R. J., & Snyder, R. D. (2011). Forecasting time series with complex seasonal patterns using exponential smoothing. *Journal of the American Statistical Association*, *106*(496), 1513-1527. https://doi.org/https://doi.org/10.1198/jasa.2011.tm09771
[31]. De Menezes, L. M., Bunn, D. W., & Taylor, J. W. (2000). Review of guidelines for the use of combined forecasts. *International Journal of Forecasting*, *16*(1), 47-66. https://doi.org/https://doi.org/10.1016/S0169-2070(99)00054-7
[32]. Devaine, M., Gaillard, P., Goude, Y., & Stoltz, G. (2013). Forecasting electricity consumption by aggregating specialized experts. *Machine Learning*, *90*(2), 231-260. https://doi.org/https://doi.org/10.1007/s10994-012-5313-5
[33]. Diebold, F. X., & Mariano, R. S. (1995). Comparing predictive accuracy. *Journal of Business & Economic Statistics*, *13*(3), 253-263. https://doi.org/https://doi.org/10.1080/07350015.1995.10524599
[34]. Durbin, J., & Koopman, S. J. (1997). Monte Carlo maximum likelihood estimation for non-Gaussian state space models. *Biometrika*, *84*(2), 265-277. https://doi.org/https://doi.org/10.1093/biomet/84.2.265
[35]. Engle, R. F. (1982). Autoregressive conditional heteroskedasticity with estimates of the variance of United Kingdom inflation. *Econometrica*, *50*(4), 987-1007. https://doi.org/https://doi.org/10.2307/1912773
[36]. Engle, R. F., & Manganelli, S. (2004). CAViaR: Conditional autoregressive value at risk by regression quantiles. *Journal of Business & Economic Statistics*, *22*(4), 367-381. https://doi.org/https://doi.org/10.1198/073500104000000370
[37]. Findley, D. F., Monsell, B. C., Bell, W. R., Otto, M. C., & Chen, B.-C. (1998). New capabilities and methods of the X-12-ARIMA seasonal-adjustment program. *Journal of Business & Economic Statistics*, *16*(2), 127-152. https://doi.org/https://doi.org/10.1080/07350015.1998.10524743
[38]. Fischer, T., & Krauss, C. (2018). Deep learning with long short-term memory networks for financial market predictions. *European Journal of Operational Research*, *270*(2), 654-669. https://doi.org/https://doi.org/10.1016/j.ejor.2017.11.054
[39]. Fissler, T., & Ziegel, J. F. (2016). Higher order elicitability and Osband's principle. *Annals of Statistics*, *44*(4), 1680-1707. https://doi.org/https://doi.org/10.1214/16-AOS1439
[40]. Friedman, J. H. (2002). Stochastic gradient boosting. *Computational Statistics & Data Analysis*, *38*(4), 367-378. https://doi.org/https://doi.org/10.1214/aos/1056562461
[41]. Fryzlewicz, P. (2014). Wild binary segmentation for multiple change-point detection. *Annals of Statistics*, *42*(6), 2243-2281. https://doi.org/https://doi.org/10.1214/14-AOS1221
[42]. Giacomini, R., & White, H. (2006). Tests of conditional predictive ability. *Econometrica*, *74*(6), 1545-1578. https://doi.org/https://doi.org/10.1111/j.1468-0262.2006.00718.x
[43]. Giot, P., & Laurent, S. (2003). Value-at-risk for long and short trading positions. *Journal of Applied Econometrics*, *18*(6), 641-663. https://doi.org/https://doi.org/10.1002/jae.710
[44]. Gneiting, T., & Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, *102*(477), 359-378. https://doi.org/https://doi.org/10.1198/016214506000001437
[45]. Golam Qibria, L., & Takbir Hossen, S. (2023). Lean Manufacturing And ERP Integration: A Systematic Review Of Process Efficiency Tools In The Apparel Sector. *American Journal of Scholarly Research and Innovation*, *2*(01), 104-129. https://doi.org/10.63125/mx7j4p06
[46]. Granger, C. W. J., & Ramanathan, R. (1984). Improved methods of combining forecasts. *Journal of Forecasting*, *3*(2), 197-204. https://doi.org/https://doi.org/10.1002/for.3980030207
[47]. Gu, S., Kelly, B., & Xiu, D. (2020). Empirical asset pricing via machine learning. *Review of Financial Studies*, *33*(5), 2223-2273. https://doi.org/https://doi.org/10.1093/rfs/hhaa009
[48]. Hamilton, J. D. (1989). A new approach to the economic analysis of nonstationary time series and the business cycle. *Econometrica*, *57*(2), 357-384. https://doi.org/https://doi.org/10.2307/1912559
[49]. Hansen, P. R., Huang, Z., & Shek, H. H. (2012). Realized GARCH: A joint model for returns and realized measures of volatility. *Journal of Applied Econometrics*, *27*(6), 877-906. https://doi.org/https://doi.org/10.1002/jae.1234
[50]. Hansen, P. R., & Lunde, A. (2005). A forecast comparison of volatility models: Does anything beat a GARCH(1,1)? *Journal of Applied Econometrics*, *20*(7), 873-889. https://doi.org/https://doi.org/10.1002/jae.800

[51]. Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, *9*(8), 1735-1780. https://doi.org/https://doi.org/10.1162/neco.1997.9.8.1735
[52]. Hoeting, J. A., Madigan, D., Raftery, A. E., & Volinsky, C. T. (1999). Bayesian model averaging: A tutorial. *Statistical Science*, *14*(4), 382-417. https://doi.org/https://doi.org/10.1214/ss/1009212519
[53]. Hoseinzade, E., & Haratizadeh, S. (2019). CNNpred: CNN-based stock market prediction using a diverse set of variables. *Expert Systems with Applications*, *129*, 273-285. https://doi.org/https://doi.org/10.1016/j.eswa.2019.03.029
[54]. Hosne Ara, M., Tonmoy, B., Mohammad, M., & Md Mostafizur, R. (2022). AI-ready data engineering pipelines: a review of medallion architecture and cloud-based integration models. *American Journal of Scholarly Research and Innovation*, *1*(01), 319-350. https://doi.org/10.63125/51kxtf08
[55]. Hyndman, R. J., & Koehler, A. B. (2006). Another look at measures of forecast accuracy. *International Journal of Forecasting*, *22*(4), 679-688. https://doi.org/https://doi.org/10.1016/j.ijforecast.2006.03.001
[56]. Hyndman, R. J., Koehler, A. B., Ord, J. K., & Snyder, R. D. (2002). A state space framework for automatic forecasting using exponential smoothing methods. *International Journal of Forecasting*, *18*(3), 439-454. https://doi.org/https://doi.org/10.1016/S0169-2070(01)00110-8
[57]. Istiaque, M., Dipon Das, R., Hasan, A., Samia, A., & Sayer Bin, S. (2023). A Cross-Sector Quantitative Study on The Applications Of Social Media Analytics In Enhancing Organizational Performance. *American Journal of Scholarly Research and Innovation*, *2*(02), 274-302. https://doi.org/10.63125/d8ree044
[58]. Istiaque, M., Dipon Das, R., Hasan, A., Samia, A., & Sayer Bin, S. (2024). Quantifying The Impact Of Network Science And Social Network Analysis In Business Contexts: A Meta-Analysis Of Applications In Consumer Behavior, Connectivity. *International Journal of Scientific Interdisciplinary Research*, *5*(2), 58-89. https://doi.org/10.63125/vgkwe938
[59]. Jahid, M. K. A. S. R. (2022). Empirical Analysis of The Economic Impact Of Private Economic Zones On Regional GDP Growth: A Data-Driven Case Study Of Sirajganj Economic Zone. *American Journal of Scholarly Research and Innovation*, *1*(02), 01-29. https://doi.org/10.63125/je9w1c40
[60]. Karim, F., Majumdar, S., Darabi, H., & Harford, S. (2019). Multivariate LSTM-FCNs for time series classification. *Neural Networks*, *116*, 237-245. https://doi.org/https://doi.org/10.1016/j.neunet.2019.04.014
[61]. Khashei, M., & Bijari, M. (2011). A novel hybridization of artificial neural networks and ARIMA models for time series forecasting. *Applied Soft Computing*, *11*(2), 2664-2675. https://doi.org/https://doi.org/10.1016/j.asoc.2010.10.015
[62]. Khashei, M., & Bijari, M. (2012). A novel hybridization of artificial neural networks and ARIMA models for time series forecasting. *Knowledge-Based Systems*, *27*, 1-8. https://doi.org/https://doi.org/10.1016/j.knosys.2012.02.016
[63]. Killick, R., Fearnhead, P., & Eckley, I. A. (2012). Optimal detection of changepoints with a linear computational cost. *Journal of the American Statistical Association*, *107*(500), 1590-1598. https://doi.org/https://doi.org/10.1080/01621459.2012.737745
[64]. Koenker, R., & Bassett, G. (1978). Regression quantiles. *Econometrica*, *46*(1), 33-50. https://doi.org/https://doi.org/10.2307/1913643
[65]. Koop, G., & Korobilis, D. (2012). Forecasting inflation using dynamic model averaging. *Macroeconomic Dynamics*, *16*(S2), 268-298. https://doi.org/https://doi.org/10.1017/S1365100511000462
[66]. Kourentzes, N., Barrow, D. K., & Crone, S. F. (2014). Neural network ensemble operators for time series forecasting. *Expert Systems with Applications*, *41*(9), 4235-4244. https://doi.org/https://doi.org/10.1016/j.eswa.2013.12.011
[67]. Krauss, C., Do, X. A., & Huck, N. (2017). Deep neural networks, gradient-boosted trees, random forests: Statistical arbitrage on the S&P 500. *European Journal of Operational Research*, *259*(2), 689-702. https://doi.org/https://doi.org/10.1016/j.ejor.2016.10.031
[68]. Kupiec, P. H. (1995). Techniques for verifying the accuracy of risk measurement models. *Journal of Derivatives*, *3*(2), 73-84. https://doi.org/https://doi.org/10.3905/jod.1995.407942
[69]. Kutub Uddin, A., Md Mostafizur, R., Afrin Binta, H., & Maniruzzaman, B. (2022). Forecasting Future Investment Value with Machine Learning, Neural Networks, And Ensemble Learning: A Meta-Analytic Study. *Review of Applied Science and Technology*, *1*(02), 01-25. https://doi.org/10.63125/edxgjg56
[70]. Lai, G., Chang, W.-C., Yang, Y., & Liu, H. (2018). Modeling long- and short-term temporal patterns with deep neural networks. Proceedings of the 41st International ACM SIGIR Conference on Research and Development in Information Retrieval,
[71]. Lessmann, S., Baesens, B., Seow, H.-V., & Thomas, L. C. (2015). Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research. *Decision Support Systems*, *62*, 106-120. https://doi.org/https://doi.org/10.1016/j.dss.2014.10.007
[72]. 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*(1), 56-67. https://doi.org/https://doi.org/10.1038/s42256-019-0138-9

[73]. Makridakis, S., Spiliotis, E., & Assimakopoulos, V. (2018). The M4 competition: Results, findings, conclusion and way forward. *International Journal of Forecasting*, *34*(4), 802-808. https://doi.org/https://doi.org/10.1016/j.ijforecast.2018.06.001

[74]. Makridakis, S., Spiliotis, E., & Petropoulos, F. (2022). M5 accuracy competition: Results, findings, and conclusions. *International Journal of Forecasting*, *38*(4), 1346-1364. https://doi.org/https://doi.org/10.1016/j.ijforecast.2021.08.006

[75]. Mansura Akter, E. (2023). Applications Of Allele-Specific PCR In Early Detection of Hereditary Disorders: A Systematic Review Of Techniques And Outcomes. *Review of Applied Science and Technology*, *2*(03), 1-26. https://doi.org/10.63125/n4h7t156

[76]. Mansura Akter, E., & Md Abdul Ahad, M. (2022). In Silico drug repurposing for inflammatory diseases: a systematic review of molecular docking and virtual screening studies. *American Journal of Advanced Technology and Engineering Solutions*, *2*(04), 35-64. https://doi.org/10.63125/j1hbts51

[77]. Mansura Akter, E., & Shaiful, M. (2024). A systematic review of SNP polymorphism studies in South Asian populations: implications for diabetes and autoimmune disorders. *American Journal of Scholarly Research and Innovation*, *3*(01), 20-51. https://doi.org/10.63125/8nvxcb96

[78]. McNeil, A. J., & Frey, R. (2000). Estimation of tail-related risk measures for heteroscedastic financial time series: An extreme value approach. *Journal of Empirical Finance*, *7*(3–4), 271-300. https://doi.org/https://doi.org/10.1016/S0927-5398(00)00012-8

[79]. Md Arifur, R., & Sheratun Noor, J. (2022). A Systematic Literature Review of User-Centric Design In Digital Business Systems: Enhancing Accessibility, Adoption, And Organizational Impact. *Review of Applied Science and Technology*, *1*(04), 01-25. https://doi.org/10.63125/ndjkpm77

[80]. Md Ashiqur, R., Md Hasan, Z., & Afrin Binta, H. (2025). A meta-analysis of ERP and CRM integration tools in business process optimization. *ASRC Procedia: Global Perspectives in Science and Scholarship*, *1*(01), 278-312. https://doi.org/10.63125/yah70173

[81]. Md Hasan, Z. (2025). AI-Driven business analytics for financial forecasting: a systematic review of decision support models in SMES. *Review of Applied Science and Technology*, *4*(02), 86-117. https://doi.org/10.63125/gjrpv442

[82]. Md Hasan, Z., Mohammad, M., & Md Nur Hasan, M. (2024). Business Intelligence Systems In Finance And Accounting: A Review Of Real-Time Dashboarding Using Power BI & Tableau. *American Journal of Scholarly Research and Innovation*, *3*(02), 52-79. https://doi.org/10.63125/fy4w7w04

[83]. Md Hasan, Z., Sheratun Noor, J., & Md. Zafor, I. (2023). Strategic role of business analysts in digital transformation tools, roles, and enterprise outcomes. *American Journal of Scholarly Research and Innovation*, *2*(02), 246-273. https://doi.org/10.63125/rc45z918

[84]. Md Mahamudur Rahaman, S. (2022). Electrical And Mechanical Troubleshooting in Medical And Diagnostic Device Manufacturing: A Systematic Review Of Industry Safety And Performance Protocols. *American Journal of Scholarly Research and Innovation*, *1*(01), 295-318. https://doi.org/10.63125/d68y3590

[85]. Md Masud, K., Mohammad, M., & Sazzad, I. (2023). Mathematics For Finance: A Review of Quantitative Methods In Loan Portfolio Optimization. *International Journal of Scientific Interdisciplinary Research*, *4*(3), 01-29. https://doi.org/10.63125/j43ayz68

[86]. Md Nur Hasan, M., Md Musfiqur, R., & Debashish, G. (2022). Strategic Decision-Making in Digital Retail Supply Chains: Harnessing AI-Driven Business Intelligence From Customer Data. *Review of Applied Science and Technology*, *1*(03), 01-31. https://doi.org/10.63125/6a7rpy62

[87]. Md Sultan, M., Proches Nolasco, M., & Md. Torikul, I. (2023). Multi-Material Additive Manufacturing For Integrated Electromechanical Systems. *American Journal of Interdisciplinary Studies*, *4*(04), 52-79. https://doi.org/10.63125/y2ybrx17

[88]. Md Sultan, M., Proches Nolasco, M., & Vicent Opiyo, N. (2025). A Comprehensive Analysis Of Non-Planar Toolpath Optimization In Multi-Axis 3D Printing: Evaluating The Efficiency Of Curved Layer Slicing Strategies. *Review of Applied Science and Technology*, *4*(02), 274-308. https://doi.org/10.63125/5fdxa722

[89]. Md Takbir Hossen, S., Ishtiaque, A., & Md Atiqur, R. (2023). AI-Based Smart Textile Wearables For Remote Health Surveillance And Critical Emergency Alerts: A Systematic Literature Review. *American Journal of Scholarly Research and Innovation*, *2*(02), 1-29. https://doi.org/10.63125/ceqapd08

[90]. Md Takbir Hossen, S., & Md Atiqur, R. (2022). Advancements In 3d Printing Techniques For Polymer Fiber-Reinforced Textile Composites: A Systematic Literature Review. *American Journal of Interdisciplinary Studies*, *3*(04), 32-60. https://doi.org/10.63125/s4r5m391

[91]. Md Tawfiqul, I. (2023). A Quantitative Assessment Of Secure Neural Network Architectures For Fault Detection In Industrial Control Systems. *Review of Applied Science and Technology*, *2*(04), 01-24. https://doi.org/10.63125/3m7gbs97

[92]. Md Tawfiqul, I., Meherun, N., Mahin, K., & Mahmudur Rahman, M. (2022). Systematic Review of Cybersecurity Threats In IOT Devices Focusing On Risk Vectors Vulnerabilities And Mitigation Strategies.

American Journal of Scholarly Research and Innovation, *1*(01), 108-136. https://doi.org/10.63125/wh17mf19

[93]. Md Tawfiqul, I., Sabbir, A., Md Anikur, R., & Md Arifur, R. (2024). Neural Network–Based Risk Prediction And Simulation Framework For Medical IOT Cybersecurity: An Engineering Management Model For Smart Hospitals. *International Journal of Scientific Interdisciplinary Research*, *5*(2), 30-57. https://doi.org/10.63125/g0mvct35

[94]. Md.Kamrul, K., & Md Omar, F. (2022). Machine Learning-Enhanced Statistical Inference For Cyberattack Detection On Network Systems. *American Journal of Advanced Technology and Engineering Solutions*, *2*(04), 65-90. https://doi.org/10.63125/sw7jzx60

[95]. Meinshausen, N. (2006). Quantile regression forests. *Journal of Machine Learning Research*, *7*, 983-999. https://doi.org/https://doi.org/10.1007/s10994-007-0072-4

[96]. Montero-Manso, P., & Hyndman, R. J. (2020). Principles and algorithms for forecasting groups of time series: Locality and globality. *International Journal of Forecasting*, *36*(1), 45-59. https://doi.org/https://doi.org/10.1016/j.ijforecast.2019.11.006

[97]. Mst Shamima, A., Niger, S., Md Atiqur Rahman, K., & Mohammad, M. (2023). Business Intelligence-Driven Healthcare: Integrating Big Data And Machine Learning For Strategic Cost Reduction And Quality Care Delivery. *American Journal of Interdisciplinary Studies*, *4*(02), 01-28. https://doi.org/10.63125/crv1xp27

[98]. Mubashir, I., & Abdul, R. (2022). Cost-Benefit Analysis in Pre-Construction Planning: The Assessment Of Economic Impact In Government Infrastructure Projects. *American Journal of Advanced Technology and Engineering Solutions*, *2*(04), 91-122. https://doi.org/10.63125/kjwd5e33

[99]. Natekin, A., & Knoll, A. (2013). Gradient boosting machines, a tutorial. *Frontiers in Neurorobotics*, *7*(21), 1-21. https://doi.org/https://doi.org/10.3389/fnbot.2013.00021

[100]. Nelson, D. M. Q., Pereira, A. C. M., & de Oliveira, R. A. (2017). Stock market's price movement prediction with LSTM neural networks. 2017 International Joint Conference on Neural Networks (IJCNN),

[101]. Niculescu-Mizil, A., & Caruana, R. (2005). Predicting good probabilities with supervised learning. Proceedings of the 22nd International Conference on Machine Learning,

[102]. Pai, P.-F., & Lin, C.-S. (2005). A hybrid ARIMA and support vector machines model in stock price forecasting. *Omega*, *33*(6), 497-505. https://doi.org/https://doi.org/10.1016/j.omega.2004.07.024

[103]. Patton, A. J., Ziegel, J. F., & Chen, R. (2019). Dynamic semiparametric models for expected shortfall (and value-at-risk). *Journal of Econometrics*, *211*(2), 388-413. https://doi.org/https://doi.org/10.1016/j.jeconom.2018.10.008

[104]. Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V., & Gulin, A. (2018). CatBoost: Unbiased boosting with categorical features. Advances in Neural Information Processing Systems,

[105]. Qin, Y., Song, D., Chen, H., Cheng, W., Jiang, G., & Cottrell, G. W. (2017). A dual-stage attention-based recurrent neural network for time series prediction. Proceedings of the 26th International Joint Conference on Artificial Intelligence (IJCAI),

[106]. Reduanul, H., & Mohammad Shoeb, A. (2022). Advancing AI in Marketing Through Cross Border Integration Ethical Considerations And Policy Implications. *American Journal of Scholarly Research and Innovation*, *1*(01), 351-379. https://doi.org/10.63125/d1xg3784

[107]. Rezwanul Ashraf, R., & Hosne Ara, M. (2023). Visual communication in industrial safety systems: a review of UI/UX design for risk alerts and warnings. *American Journal of Scholarly Research and Innovation*, *2*(02), 217-245. https://doi.org/10.63125/wbv4z521

[108]. Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?": Explaining the predictions of any classifier. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining,

[109]. Salinas, D., Flunkert, V., Gasthaus, J., & Januschowski, T. (2020). DeepAR: Probabilistic forecasting with autoregressive recurrent networks. *International Journal of Forecasting*, *36*(4), 1181-1191. https://doi.org/https://doi.org/10.1016/j.ijforecast.2019.07.001

[110]. Sanjai, V., Sanath Kumar, C., Maniruzzaman, B., & Farhana Zaman, R. (2023). Integrating Artificial Intelligence in Strategic Business Decision-Making: A Systematic Review Of Predictive Models. *International Journal of Scientific Interdisciplinary Research*, *4*(1), 01-26. https://doi.org/10.63125/s5skge53

[111]. Sanjai, V., Sanath Kumar, C., Sadia, Z., & Rony, S. (2025). AI And Quantum Computing For Carbon-Neutral Supply Chains: A Systematic Review Of Innovations. *American Journal of Interdisciplinary Studies*, *6*(1), 40-75. https://doi.org/10.63125/nrdx7d32

[112]. Sazzad, I., & Md Nazrul Islam, K. (2022). Project impact assessment frameworks in nonprofit development: a review of case studies from south asia. *American Journal of Scholarly Research and Innovation*, *1*(01), 270-294. https://doi.org/10.63125/eeja0t77

[113]. Scott, S. L., & Varian, H. R. (2014). Predicting the present with Bayesian structural time series. *International Journal of Mathematical Modelling and Numerical Optimisation*, *5*(1/2), 4-23. https://doi.org/https://doi.org/10.1504/IJMMNO.2014.059942

[114]. 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/https://doi.org/10.1016/j.asoc.2020.106181

[115]. Sheratun Noor, J., & Momena, A. (2022). Assessment Of Data-Driven Vendor Performance Evaluation in Retail Supply Chains: Analyzing Metrics, Scorecards, And Contract Management Tools. *American Journal of Interdisciplinary Studies*, *3*(02), 36-61. https://doi.org/10.63125/0s7t1y90

[116]. Sill, J., Tarlow, D., Bousquet, O., & Lal, T. N. (2009). Feature-weighted linear stacking. *arXiv preprint*. https://doi.org/https://doi.org/10.48550/arXiv.0911.0460

[117]. Sirignano, J., & Cont, R. (2019). Universal features of price formation in financial markets: Perspectives from deep learning. *Quantitative Finance*, *19*(9), 1449-1459. https://doi.org/https://doi.org/10.1080/14697688.2019.1622295

[118]. Smyl, S. (2020). A hybrid method of exponential smoothing and recurrent neural networks for time series forecasting. *International Journal of Forecasting*, *36*(1), 75-85. https://doi.org/https://doi.org/10.1016/j.ijforecast.2019.03.012

[119]. Subrato, S., & Md, N. (2024). The role of perceived environmental responsibility in artificial intelligence-enabled risk management and sustainable decision-making. *American Journal of Advanced Technology and Engineering Solutions*, *4*(04), 33-56. https://doi.org/10.63125/7tjw3767

[120]. Sundararajan, M., Taly, A., & Yan, Q. (2017). Axiomatic attribution for deep networks. *arXiv preprint*. https://doi.org/https://doi.org/10.48550/arXiv.1703.01365

[121]. Tahmina Akter, R., Debashish, G., Md Soyeb, R., & Abdullah Al, M. (2023). A Systematic Review of AI-Enhanced Decision Support Tools in Information Systems: Strategic Applications In Service-Oriented Enterprises And Enterprise Planning. *Review of Applied Science and Technology*, *2*(01), 26-52. https://doi.org/10.63125/73djw422

[122]. Taylor, J. W. (2003). Short-term electricity demand forecasting using double seasonal exponential smoothing. *Journal of the Operational Research Society*, *54*(8), 799-805. https://doi.org/https://doi.org/10.1057/palgrave.jors.2601589

[123]. Taylor, J. W. (2008). Estimating value at risk and expected shortfall using expectiles. *Journal of Financial Econometrics*, *6*(2), 231-252. https://doi.org/https://doi.org/10.1093/jjfinec/nbn001

[124]. Taylor, J. W. (2010). Triple seasonal methods for short-term electricity demand forecasting. *European Journal of Operational Research*, *204*(1), 139-152. https://doi.org/https://doi.org/10.1016/j.ejor.2009.10.003

[125]. Taylor, J. W. (2019). Forecasting value at risk and expected shortfall using a semiparametric approach based on the asymmetric Laplace distribution. *Journal of Business & Economic Statistics*, *37*(1), 121-133. https://doi.org/https://doi.org/10.1080/07350015.2017.1281815

[126]. Taylor, S. J., & Letham, B. (2018). Forecasting at scale. *The American Statistician*, *72*(1), 37-45. https://doi.org/https://doi.org/10.1080/00031305.2017.1380080

[127]. Timmermann, A. (2006). Forecast combinations. *Journal of Business & Economic Statistics*, *24*(1), 129-141. https://doi.org/https://doi.org/10.1198/073500106X115016

[128]. Ting, K. M., & Witten, I. H. (1999). Issues in stacked generalization. *Journal of Artificial Intelligence Research*, *10*, 271-289. https://doi.org/https://doi.org/10.1613/jair.614

[129]. van der Laan, M. J., Polley, E. C., & Hubbard, A. E. (2007). Super Learner. *Statistical Applications in Genetics and Molecular Biology*, *6*(1), Article 25. https://doi.org/https://doi.org/10.2202/1544-6115.1309

[130]. Wang, J.-J., Wang, J.-Z., Zhang, Z.-G., & Guo, S.-P. (2012). Stock index forecasting based on a hybrid model. *Expert Systems with Applications*, *39*(5), 8891-8901. https://doi.org/https://doi.org/10.1016/j.eswa.2012.01.150

[131]. Wolpert, D. H. (1992). Stacked generalization. *Neural Networks*, *5*(2), 241-259. https://doi.org/https://doi.org/10.1016/S0893-6080(05)80023-1

[132]. Yang, Y. (2004). Aggregating regression procedures to improve performance. *Bernoulli*, *10*(1), 25-47. https://doi.org/https://doi.org/10.3150/bj/1077544606

[133]. Zadrozny, B., & Elkan, C. (2002). Transforming classifier scores into accurate multiclass probability estimates. Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining,

[134]. Zhang, G. P. (2003). Time series forecasting using a hybrid ARIMA and neural network model. *Neurocomputing*, *50*, 159-175. https://doi.org/https://doi.org/10.1016/S0925-2312(01)00702-0

[135]. Zhang, Z., Zohren, S., & Roberts, S. (2019). DeepLOB: Deep convolutional neural networks for limit order books. *IEEE Transactions on Signal Processing*, *67*(11), 3001-3012. https://doi.org/https://doi.org/10.1109/TSP.2019.2907260