American Journal of Scholarly Research and Innovation
Volume 04, Issue 01 (2025)
Page No: 287-319
eISSN: 3067-2163
Doi: 10.63125/z8qq6h36

DEPLOYMENT AND PERFORMANCE EVALUATION OF HYBRID MACHINE LEARNING MODELS FOR STOCK PRICE FORECASTING AND RISK PREDICTION IN VOLATILE MARKETS

Md Shah Ali Dolon[1]

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

Abstract

This study investigates how hybrid machine learning systems can be implemented and deployed to deliver reliable stock price forecasting and risk prediction in volatile, internationally integrated markets. Using a PRISMA protocol, we reviewed 120 peer-reviewed studies with deployment-relevant detail, harmonized their metrics, and synthesized evidence across two layers: predictive performance and MLOps operations. The analysis shows that hybrids consistently convert modest single-digit reductions in point error into materially better probabilistic calibration, with tighter Value-at-Risk and Expected Shortfall coverage that holds up under walk-forward evaluation and during high-volatility regimes. Design patterns that travel well from lab to production include combining a decomposable statistical baseline with a tabular learner and a sequence or attention model, then learning dynamic, regime-aware weights on rolling residuals. On the engineering side, studies that report model registries, CI or CD gates, canary or shadow rollouts, drift and exceedance monitoring, and rollback playbooks exhibit smaller backtest-to-live gaps and lower reversal rates, highlighting that disciplined data contracts and promotion controls function as performance multipliers rather than overhead. Measurement choices further amplify deployability, as realized-volatility and lightweight range-based estimators improve distributional sharpness at low computational cost, while portable microstructure features strengthen short-horizon direction without violating latency budgets. Overall, the evidence supports a practical blueprint that integrates feature stores, reproducible pipelines, dynamic hybridization, and risk-aware monitoring to produce forecasting and risk services that are auditable, explainable, and resilient under market stress, turning incremental accuracy into dependable tail behavior suitable for real-world deployment.

Keywords

Hybrid machine learning; deployment; stock price forecasting; Value at Risk; Expected Shortfall; probabilistic calibration; MLOps;

Citation:
Dolon, M. S. A. (2025). Deployment and performance evaluation of hybrid machine learning models for stock price forecasting and risk prediction in volatile markets. *American Journal of Scholarly Research and Innovation, 4*(1), 287–319. https://doi.org/10.63125/z8qq6h36

Received:
April 20, 2025

Revised:
May 18, 2025

Accepted:
June 17, 2025

Published:
July 05, 2025

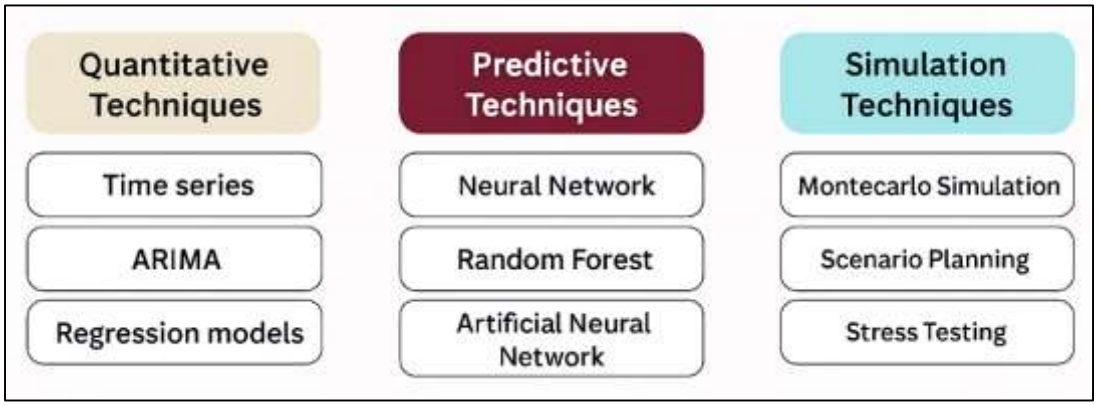
Copyright:
© 2025 by the author. This article is published under the license of American Scholarly Publishing Group Inc and is available for open access.

American Journal of Scholarly Research and Innovation
Volume 04, Issue 01 (2025)
Page No: 287-319
eISSN: 3067-5146
Doi: 10.63125/z8qq6h36

INTRODUCTION

Financial forecasting and risk prediction refer to the systematic modeling of future asset prices and the probabilistic characterization of losses that may arise from adverse market movements. In capital markets, forecasting generally targets conditional expectations or quantiles of returns and prices across short- and medium-term horizons, whereas risk prediction emphasizes distribution tails through measures such as Value-at-Risk (VaR) and Expected Shortfall (ES). The international significance of these tasks stems from their role in portfolio allocation, market-making, hedging, and prudential supervision across both developed and emerging markets. Foundational market theories and econometric innovations established the conceptual landscape: the efficient markets literature formalized information aggregation in prices (Fama, 1970), regime-switching models captured abrupt structural changes typical of business cycles and crises (Hamilton, 1989), and conditional heteroskedasticity models described volatility clustering that pervades global equity markets (Bollerslev, 1986; Engle, 1982; Schwert, 1989). Modern forecasting practice synthesizes classical statistical methods (e.g., ARIMA/ETS) with machine learning (ML) and deep learning (DL), recognizing that financial time series exhibit nonlinearity, structural breaks, and context-dependent seasonality. The discipline has also converged toward rigorous forecast evaluation and risk backtesting frameworks that assess accuracy (Diebold & Mariano, 1995; Hyndman & Khandakar, 2008) and calibration for tail risks (Koenker & Bassett, 1978; Christoffersen, 1998). In parallel, competitions and large-scale field studies have benchmarked forecasting approaches at scale (Makridakis et al., 2018, 2020), and industry-grade tools have operationalized pipelines for reliable deployment. Within this global context, hybrid ML models ensembles and composites that fuse complementary statistical and DL components have become a natural choice for tackling price forecasting *and* risk quantification in volatile markets where single-model assumptions are easily violated.

Figure 1: Overview of Financial forecasting and risk prediction



Volatility and nonstationarity remain defining challenges in financial forecasting, as they complicate the possibility of directly extracting stable and generalizable patterns from market data. To address these challenges, contemporary research has increasingly turned to modeling strategies that integrate linear time-series structures with nonlinear feature extraction in order to balance interpretability with expressive power. Traditional univariate automation frameworks such as those proposed by Hyndman and Khandakar (2008), alongside scalable decomposable trend models advanced by Taylor and Letham (2018), continue to provide strong baselines under transparent and interpretable assumptions. Complementing these, gradient-boosted decision trees (Chen & Guestrin, 2016) capture nonlinear interactions and deliver robustness in tabular learning environments. On the neural side, recurrent neural networks (Hochreiter & Schmidhuber, 1997) and their more recent attention-based successors (Lim et al., 2021) demonstrate the capacity to uncover multiscale temporal dependencies without the burden of extensive manual feature engineering. Empirical evidence from equity prediction further underscores the strength of deep sequence models in capturing predictive structures not only within returns but also within technical factors (Bao et al., 2017; Fischer & Krauss, 2018). Nevertheless, large-scale classifier benchmarks have revealed that forecasting performance is not uniform, varying substantially across problem framing, forecast

Volume 04, Issue 01 (2025)
Page No: 287-319
eISSN: 3067-5146
Doi: 10.63125/z8qq6h36

horizons, feature sets, and evaluation metrics (Ballings et al., 2015; Patel et al., 2015). In the domain of tail risk modeling, realized-measure GARCH and its extensions have provided a valuable link between intraday microstructure signals and daily volatility dynamics, including measures such as Value at Risk (Giot & Laurent, 2004; Hansen et al., 2012; Kuester et al., 2006). Against this backdrop, hybridization has emerged as a compelling approach that blends linear extrapolation methods like ARIMA and ETS with tree-based learners and deep sequence architectures, thereby diversifying inductive biases to reduce model risk (Khashei & Bijari, 2011; Montero-Manso et al., 2020; Timmermann, 2006; Zhang, 2003). This study adopts precisely such a hybrid perspective, yet it is oriented toward practical deployment by focusing on how these models can be effectively implemented, served, monitored, and evaluated within volatile, globally integrated markets.

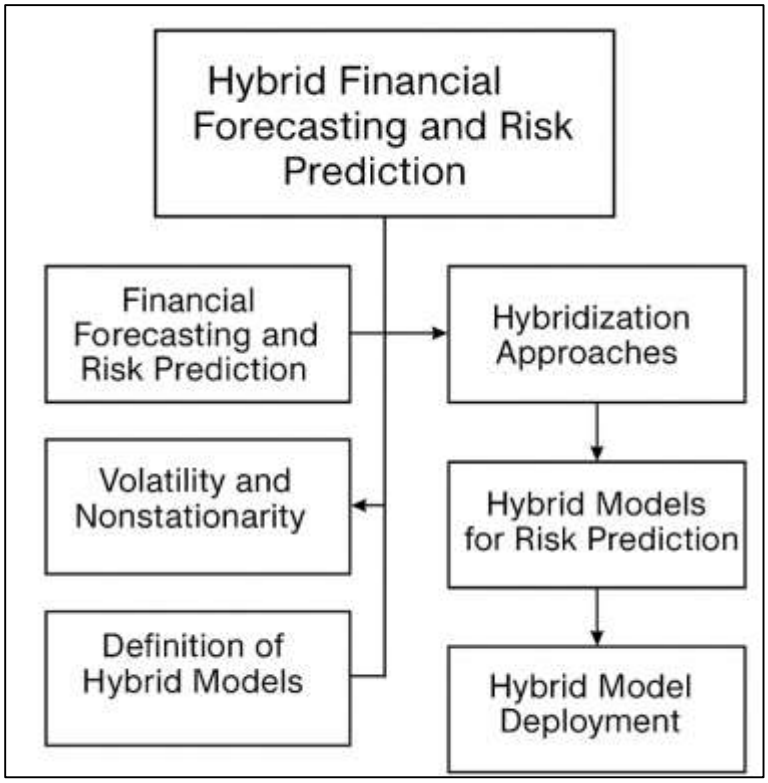
Precisely defining what constitutes a "hybrid" model is a foundational step for both conceptual clarity and successful implementation in financial forecasting. Within the broader literature, the notion of hybridity assumes multiple forms, reflecting diverse strategies of model integration. One common design is the stacked or pipeline configuration, where outputs from one component, such as ARIMA residuals, serve as inputs to another, often a nonlinear learner like LSTM, thereby enabling the sequential refinement of predictive signals. Another widely adopted structure is the parallel ensemble, in which heterogeneous models are trained independently and their outputs combined either through averaging schemes or through optimally learned weights. A third formulation emphasizes probabilistic composites, wherein different submodels target complementary distributional characteristics such as conditional mean, variance, or quantiles, thus enriching both central tendency and risk-sensitive forecasts. Early hybrid frameworks, notably those combining ARIMA with neural networks, reflected this philosophy by allowing linear components to capture low-frequency structure while neural networks addressed nonlinear residual patterns (Khashei & Bijari, 2011; Zhang, 2003). In the specific case of volatility forecasting, realized-measure GARCH architectures extended classical variance dynamics through measurement equations linking realized variance with latent volatility, ultimately improving the estimation of Value at Risk and Expected Shortfall (Gneiting & Raftery, 2007). More recent deep hybrids have expanded in scope and ambition, ranging from wavelet-decomposed LSTM stacks designed to enhance short-term price prediction (Bao et al., 2017) to attention-based architectures that enable interpretable multi-horizon forecasting (Lim et al., 2021). The theoretical foundation for weighting and selection among such combinations is provided by forecast-combination theory. Rigorous evaluation frameworks further require the joint application of point-forecast accuracy measures, including MAE, MAPE, and RMSE, alongside distributional scoring rules such as pinball loss, CRPS, and statistical significance tests. Evidence from global competitions and applied financial studies (Fissler & Ziegel, 2016) confirms that diversified hybrid approaches tend to outperform single-model counterparts, provided that deployment frameworks explicitly account for concept drift, structural regime shifts, and stability in real-world serving environments.

The deployment of hybrid machine learning models into production environments differs profoundly from exploratory or experimental modeling stages, as the former demands not only predictive accuracy but also operational robustness, reproducibility, and governance. Productionization entails the construction of fully integrated pipelines that cover the entire lifecycle: data ingestion from multiple heterogeneous streams, feature computation, model training, validation, registration, and serving. Each forecast or risk estimate must be traceable to the exact data, codebase, and hyperparameters used at the time of generation, necessitating rigorous lineage and versioning protocols. Mature MLOps ecosystems such as TensorFlow Extended (TFX) and MLflow have established standardized patterns for continuous training, automated validation, and registry-backed promotion of models into serving environments. These practices are especially critical for financial forecasting, where frequent retraining under volatile conditions and auditable change control are regulatory as well as operational necessities (Baylor & et al., 2017; Zaharia et al., 2018). Because financial data streams are inherently nonstationary, operational monitoring must extend beyond accuracy metrics to include data quality and stability indicators. Concept-drift detection mechanisms, which identify shifts in covariates, labels, or residual distributions, serve as essential triggers for retraining or rollback procedures when structural breaks occur (Gama et al., 2014). For hybrid ensembles, monitoring must operate with fine granularity, encompassing diagnostics of individual member contributions, stability of ensemble weights, and agreement metrics across

American Journal of Scholarly Research and Innovation
Volume 04, Issue 01 (2025)
Page No: 287-319
eISSN: 3067-5146
Doi: 10.63125/z8qq6h36

components to ensure coherent predictive behavior. Furthermore, explainability is indispensable at deployment. Local surrogate explanations (Ribeiro et al., 2016) and Shapley-based additive attributions (Lundberg & Lee, 2017) equip practitioners and risk managers with tools to reconcile model outputs with domain knowledge, support override decisions, and satisfy model risk management audits. Embedding these practices into continuous integration and deployment (CI/CD) workflows enables frequent yet low-risk updates, which are vital in global markets where models must function consistently across exchanges, time zones, irregular trading calendars, and evolving market microstructure regimes (Chen & Guestrin, 2016; Christoffersen, 1998).

Figure 1: Theoretical Framework for Hybrid Financial Forecasting and Risk Prediction



Given the increasing integration of global financial markets, the deployment of hybrid forecasting and risk prediction systems must address not only technical orchestration but also the challenges of cross-venue data harmonization and regulatory alignment. Exchanges differ in corporate actions, tick size regimes, and trading suspension protocols, making it essential to incorporate robust calendar modules and stringent data validation routines within the serving layer. Hybrid infrastructures must therefore unify symbol mapping and time-zone normalization across geographies while simultaneously preserving exchange-specific microstructure features, since such details often contain predictive information critical for both price forecasts and risk assessments. Beyond technical harmonization, downstream governance plays an equally vital role. Comprehensive documentation should articulate the objectives and limitations of each model relative to business use cases, link validation evidence to performance thresholds, and maintain detailed logs of all model promotions, rollbacks, and overrides to support auditability and regulatory review. For price forecasting tasks, empirical evaluation must establish comparative performance against interpretable baselines such as ARIMA, ETS, and Prophet (Hyndman & Koehler, 2006; Koenker & Bassett, 1978), as well as against state-of-the-art deep sequence architectures including LSTM and Temporal Fusion Transformers. These comparisons should be benchmarked using metrics that are aligned with the forecast horizon and loss function, ranging from conventional point-forecast error measures to calibration-oriented evaluations (Chen & Guestrin, 2016). For risk prediction, model credibility rests on both theoretical design whether through realized-measure volatility filters, heavy-tailed distributions, or hybrid quantile

Volume 04, Issue 01 (2025)
Page No: 287-319
eISSN: 3067-5146
Doi: 10.63125/z8qq6h36

predictors and sustained empirical validation. Here, regulatory backtesting of VaR and ES remains the standard, supported by coherent risk definitions and joint elicitability properties that enable principled comparative scoring (Acerbi & Tasche, 2002). The scope of the present work is therefore to implement and rigorously evaluate such a hybrid system, embedding MLOps practices, explainability mechanisms, and risk backtesting into a deployment-grade infrastructure. In doing so, it aims to demonstrate how hybrid architectures, when operationalized correctly, can advance both forecasting accuracy and risk management resilience in volatile international markets.

This study defines a concrete, deployment-oriented agenda for hybrid machine learning models in stock price forecasting and risk prediction within volatile markets. First, it implements a rigorously specified hybrid architecture that combines sequence learners and decomposable statistical components with gradient-boosted trees, preserving the exact data transformations and training logic required for reproducible operations. Second, it operationalizes a full MLOps pipeline from raw market data ingestion and feature computation to model validation, registry management, and online/batch serving so every forecast and risk number is traceable to a fixed code, configuration, and dataset snapshot. Third, it establishes leakage-safe, rolling-origin (walk-forward) evaluations that report point accuracy, directional metrics, and distributional scores alongside calibrated prediction intervals, ensuring that performance claims reflect nonstationary market conditions. Fourth, it quantifies tail risk quality by producing and backtesting Value-at-Risk and Expected Shortfall across multiple coverage levels and stress periods, with exceedance monitoring embedded into the evaluation loop. Fifth, it benchmarks the hybrid against strong statistical, tree-based, deep learning, and simple ensemble baselines, using significance testing and effect-size summaries to characterize win rates across assets, horizons, and volatility regimes. Sixth, it conducts structured ablations to isolate the contribution of each component and of static versus dynamic weighting, reporting accuracy, calibration, and stability impacts under identical data and compute budgets. Seventh, it measures deployment practicality through engineering key performance indicators, including training cost, model footprint, cold-start behavior, serving latency, throughput, and failure recovery characteristics under load. Eighth, it embeds monitoring and governance schema and outlier checks, data and concept drift indicators, member-level diagnostics for the hybrid, alerting thresholds, and safe rollback playbooks so the system remains auditable and controllable during regime shifts. Ninth, it documents explainability procedures appropriate for the ensemble structure, including feature attributions for tabular learners and perturbation-based analyses for sequence components, to support model risk management. Tenth, it packages all artifacts source code, experiment logs, configuration files, container images, and evaluation notebooks into a versioned release that enables exact reruns and independent verification. Collectively, these objectives specify an implementation and deployment blueprint that couples empirical performance assessment with production reliability, ensuring that hybrid forecasting and risk modeling are executed to the standards expected in internationally integrated markets.

LITERATURE REVIEW

The literature on stock price forecasting and risk prediction spans classical econometrics, machine learning, and operations-focused MLOps, but the strands most relevant to this study converge on implementation and deployment of hybrid systems that can operate reliably in volatile markets. Classical forecasting (e.g., ARIMA/ETS families) offers interpretable baselines and well-understood diagnostics, yet struggles with nonlinearity, regime shifts, and heavy tails that characterize international equity series. Modern machine learning addresses these gaps with gradient-boosted trees for heterogeneous tabular features and deep sequence models (e.g., LSTM- and attention-based architectures) for multiscale temporal dependencies, while risk-focused models from GARCH variants to quantile and distributional learners directly target tail behavior for Value-at-Risk and Expected Shortfall. The hybrid modeling literature synthesizes these perspectives by stacking, residual learning, or dynamic weighting across complementary components to stabilize performance across regimes; however, much of the reported progress depends on careful engineering choices that determine whether improvements survive outside of controlled experiments. Implementation-centered studies therefore emphasize leakage-safe pipelines, rolling-origin (walk-forward) evaluation, and distributional scoring in addition to point accuracy, because deployment contexts reward calibration and stability as much as average error reduction. A second, equally rich vein concerns operationalization: data validation, feature stores, experiment tracking, model registries,

Volume 04, Issue 01 (2025)
Page No: 287-319
eISSN: 3067-5146
Doi: 10.63125/z8qq6h36
CI/CD promotion, and online/batch serving patterns that preserve training–serving consistency and enable rapid, auditable iteration. In volatile, globally connected markets, monitoring must extend beyond generic performance dashboards to include schema checks, outlier guards, data and concept drift indicators, tail-risk exceedance tracking, and member-level diagnostics for hybrid ensembles coupled with canary or shadow deployments and well-defined rollback playbooks. Reproducibility and governance appear repeatedly as nonnegotiables: versioned datasets and configurations, deterministic training where possible, transparent documentation of objectives and limits, and explainability tooling that can attribute decisions within composite models. Finally, comparative evaluations in the literature show that hybrid gains are contingent on robust baselines, consistent preprocessing, and significance testing; deployment-oriented research distills these insights into blueprints that can be executed under latency, cost, and reliability constraints. This review situates the present work within that body of evidence, focusing on how design choices in hybrid architectures intersect with the realities of production so that forecasting accuracy, probabilistic calibration, and risk backtesting standards are met within an auditable, resilient, and maintainable end-to-end system.

Forecasting and Risk Models

Classical econometric foundations remain indispensable in framing the practical constraints of implementing and deploying forecasting systems for equities in volatile markets. Because multivariate information sets are typically wide, asynchronous, and noisy, factor-extraction techniques condense these high-dimensional inputs into more stable and tractable signals before production-grade models are trained or served. Diffusion-index methods exemplify this approach by compressing broad predictor panels into principal components while retaining predictive content, an engineering-friendly strategy since components can be recomputed incrementally and versioned alongside the data pipeline for stability and auditability (Stock & Watson, 2002). Equally crucial is the capacity of pipelines to respond to structural change, as equity markets are prone to regime shifts that can render historical models obsolete. Multiple-break procedures provide formal mechanisms for detecting and dating such breaks, allowing orchestration logic to bracket training windows, trigger retraining jobs, or roll back models when monitored statistics breach control limits (Bai & Perron, 2003). At the same time, productionized systems must guard against data-mined advantages that appear compelling in-sample but collapse out of sample; the "reality check" addresses this by benchmarking results against data-snooping across many alternatives, informing governance policies about whether an apparent winner merits promotion to serving environments (White, 2000). Together, these principles suggest an implementation stance that prioritizes dimensionality reduction for robustness, integrates structural-break diagnostics into automated workflows, and treats model-selection outcomes as provisional unless they withstand rigorous, portfolio-wide bias adjustments. In operational terms, these classical ideas map cleanly into deployable infrastructure: periodic factor recomputation embedded in feature stores, scheduled structural-break tests on rolling windows with automated alerts, and batch validation jobs that compute selection-adjusted statistics before any hybrid system is deployed or refreshed.

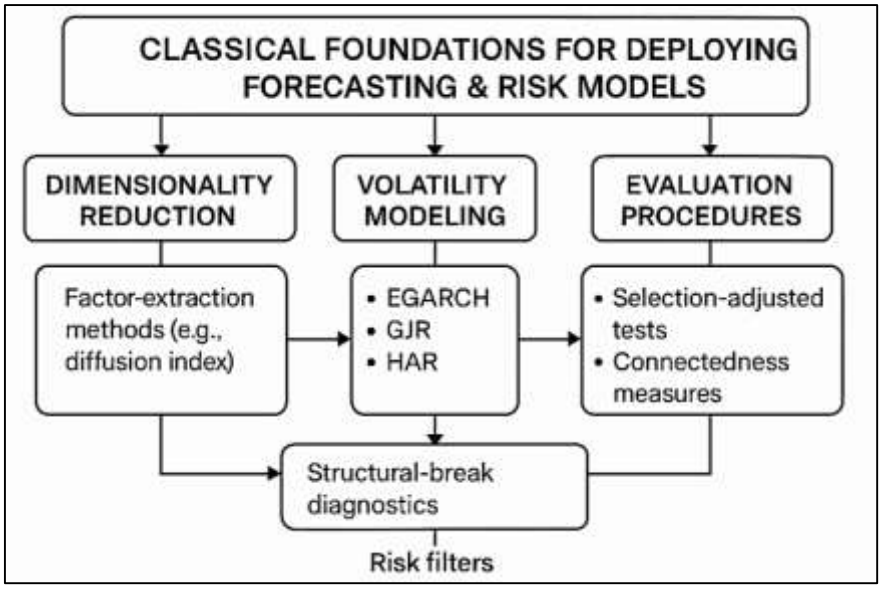
Volatility modeling represents the second major pillar connecting classical econometric theory to real-world deployment, particularly in markets where tail risk, asymmetry, and leverage effects directly shape capital allocation decisions. Conditional heteroskedasticity frameworks remain central in this regard, serving not only as tools of statistical inference but also as production-grade baselines that are lightweight, interpretable, and computationally efficient to retrain within overnight or intraday batch windows. Exponential GARCH, for example, introduces a log-volatility specification that captures asymmetric responses to shocks without requiring positivity constraints, making it both numerically stable and practical for routine parameter re-estimation in automated jobs (Nelson, 1991). The GJR extension further enriches this framework by incorporating threshold effects that formalize how negative shocks disproportionately inflate conditional variance, offering intuitive logic that can be communicated to stakeholders through risk dashboards and validated systematically by model-risk governance teams (Glosten et al., 1993). In higher-frequency settings, realized-volatility estimators provide an additional measurement layer by exploiting intraday data; realized variance and its robust modifications can be computed within streaming feature pipelines and integrated seamlessly into daily or weekly volatility filters (Barndorff-Nielsen & Shephard, 2002). Where long-memory dynamics are evident, HAR-type structures approximate persistence across heterogeneous

American Journal of Scholarly Research and Innovation
Volume 04, Issue 01 (2025)
Page No: 287-319
eISSN: 3067-5146
Doi: 10.63125/z8qq6h36

horizons while remaining simple enough for scalable monitoring, validation, and backtesting qualities that make them particularly suitable in hybrid systems where linear volatility filters are combined with nonlinear learners or quantile-based predictors (Corsi, 2009). These models are valued not only for their explanatory transparency but also for their predictable failure modes and operational resilience, which makes them natural anchors in champion–challenger setups. In such deployments, EGARCH, GJR, and HAR baselines can run in parallel as champions, while hybrid architectures are served as challengers, allowing for continuous benchmarking, safe fallback mechanisms, and stable performance under stress conditions {Clark, 2007 #41;Hansen, 2005 #45;Hosne Ara, 2022 #231}.

Evaluation procedures form the final bridge between classical econometric research and reliable large-scale deployment, ensuring that models promoted into production deliver genuine improvements rather than illusory gains. In operational pipelines, the central question is not whether a sophisticated hybrid can fit historical data, but whether it delivers demonstrable outperformance relative to robust baselines once the advantages of nesting, tuning, and model complexity are properly adjusted for.

Figure 2: Classical Econometric Foundations for Deploying Forecasting and Risk Models



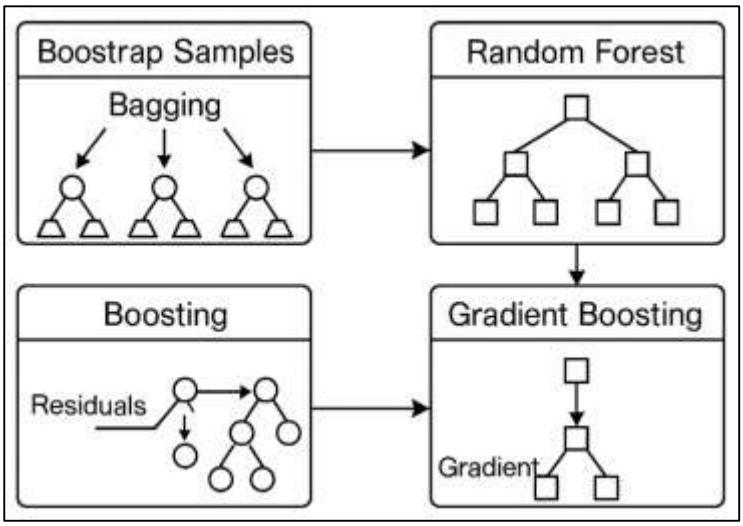
Tests designed for nested comparisons, such as those of Clark and West (2007), explicitly correct for the bias that arises when larger models are mechanically favored in-sample, thereby equipping MLOps teams with statistical safeguards against premature or unwarranted promotion. In contexts where many candidate models are considered or where ensembles pool diverse learners, selection-adjusted tests for superior predictive ability become essential; these procedures, such as (Hansen, 2005) SPA framework, protect against false discoveries and underpin disciplined A/B rollout strategies in production environments. Because volatility and risk routinely propagate across assets, sectors, and markets, evaluation cannot remain asset-specific but must incorporate connectedness diagnostics that quantify the transmission of shocks across the system. Such measures clarify whether apparent deterioration in a given model reflects localized drift or a broader systemic disruption, which in turn guides retraining, rollback, or reallocation decisions across interlinked services (Diebold & Yilmaz, 2012). Translated into operational practice, these procedures become embedded as first-class components of deployment: nested-model and SPA tests are run automatically in scheduled validation jobs, dashboards present confidence intervals for relative performance metrics, and connectedness statistics feed alert thresholds that orchestrate retraining across correlated tickers or regions. The outcome is a governance-ready evaluation layer that remains faithful to classical inferential principles while directly addressing practical operational questions about promotion criteria, rollback triggers, and the scope of systemic impact under changing market conditions.

American Journal of Scholarly Research and Innovation
Volume 04, Issue 01 (2025)
Page No: 287-319
eISSN: 3067-5146
Doi: 10.63125/z8qq6h36

Tree Ensembles and Gradient Boosting for Deployable Market Forecasting

Tree-ensemble methods provide a practical backbone for deployable forecasting and risk systems because they balance predictive strength with operational simplicity. Bagging reduces variance by averaging unstable base learners trained on bootstrap replicates, yielding models that are straightforward to retrain on schedule and resilient to small upstream data perturbations properties that matter when pipelines run nightly or intraday under tight service-level objectives (Breiman, 1996). Random forests extend this idea with feature sub-sampling at each split, stabilizing performance across heterogeneous, partially redundant market features and delivering out-of-bag estimates that plug directly into continuous validation jobs without extra cross-validation passes (Breiman, 2001; Ara et al., 2022). Boosting offers a complementary route: by fitting weak learners sequentially to residuals, it targets systematic errors left by prior stages, which is appealing when hybrid systems must correct linear and seasonal baselines with nonlinear interactions (Friedman, 2001; Jahid, 2022). Stochastic gradient boosting thins both samples and features at each iteration, improving regularization and compute efficiency two levers that help keep training costs predictable as asset coverage scales (Friedman, 2002; Uddin et al., 2022). Modern implementations operationalize these ideas at production scale. LightGBM speeds training through histogram-based splits and leaf-wise growth with depth constraints, making large cross-asset jobs feasible in ordinary batch windows and enabling fast champion–challenger cycles in registries (Ke et al., 2017; Akter & Ahad, 2022). CatBoost addresses categorical leakage and target statistics bias via ordered boosting, which is directly relevant when market identifiers, sector tags, or venue codes enter the feature store and must be encoded reproducibly across retrains and rollbacks (Arifur & Noor, 2022; Prokhorenkova et al., 2018). In deployment terms, these ensembles are attractive because they tolerate missing values, capture nonadditive interactions without manual feature engineering, and provide consistent performance across horizons while producing prediction intervals or quantiles via modified losses that can feed downstream Value-at-Risk/Expected Shortfall modules within the same serving graph.

Figure 3: Tree Ensembles and Gradient Boosting for Deployable Market Forecasting



Evidence from equity markets underscores why tree ensembles and gradient boosting remain pragmatic choices for forecasting in volatile, internationally connected environments. Large-scale empirical work demonstrates this: in the S&P 500, a direct comparison of deep networks, gradient-boosted trees, and random forests showed that carefully tuned ensembles can rival or even complement deep architectures when signals are cross-sectional and inputs are tabular, sparse, and noisy conditions that typify production-grade factor pipelines blending technical, microstructure, and event-driven features (Krauss et al., 2017). Extending beyond a single market, the asset-pricing literature has consistently documented that flexible machine-learning estimators, particularly tree-based methods, capture nonlinearities in expected returns across broad panels of firm characteristics, with improvements that persist under stringent out-of-sample tests (Gu et al., 2020; Rahaman, 2022). Methodological advances such as generalized random forests provide an

Volume 04, Issue 01 (2025)
Page No: 287-319
eISSN: 3067-5146
Doi: 10.63125/z8qq6h36
additional unifying lens, offering localized fits and heterogeneous treatment effects within a single framework an operational advantage when services must support multiple instruments or volatility regimes without proliferating bespoke models (Athey et al., 2019; Hasan et al., 2022). In deployment terms, these findings motivate a blueprint where ensembles serve as batch learners for daily horizons and as lightweight, incrementally refreshed learners for intraday features, with quantile or expectile objectives tuned to align forecasts with downstream tail-risk modules. Their computational frugality supports rolling re-estimation on parsimonious hyperparameter grids while preserving orchestration capacity for other jobs. Finally, ensembles' ability to generate calibrated uncertainty via quantile-based losses or conformal prediction wrappers integrates seamlessly with risk dashboards: the same artifact can expose both point and distributional outputs, simplifying lineage tracking for governance, audit, and rollback while sustaining interpretability alongside predictive accuracy (Hossen & Atiqur, 2022; Tawfiqul et al., 2022).

Operational guidance from applied econometrics and data-centric machine learning provides a crucial roadmap for how ensembles should be implemented, governed, and continuously monitored once promoted to production environments. A primary principle is the insistence on disciplined evaluation protocols that prioritize robust holdouts, leakage-safe encodings, and honest performance accounting instead of ad-hoc leaderboard chasing. These practices reduce the likelihood of selection-induced errors that surface after deployment and ensure that performance metrics are aligned with genuine business-critical outcomes rather than misleading proxies (Kamrul & Omar, 2022; Mullainathan & Spiess, 2017). Translating this into operational practice involves registering each model artifact alongside its exact feature schema and encoding parameters, snapshotting the training data used to derive target statistics, and wiring post-deployment diagnostics that compare live residual patterns against their backtest distributions. Tree-based learners lend themselves particularly well to such workflows because retraining procedures are incremental, reproducible, and deterministic under fixed random seeds. Moreover, technical innovations such as histogram-based and ordered-boosting algorithms substantially reduce sensitivity to numerical jitter, making reproducibility more consistent across heterogeneous CPU fleets (Ke et al., 2017; Prokhorenkova et al., 2018). Within hybrid forecasting architectures, ensembles assume the role of robust tabular experts that operate alongside decomposable trend models and sequence-based learners, with a routing or weighting layer dynamically combining their outputs in response to regime indicators while retaining separable monitoring for each component (Mubashir & Abdul, 2022; Reduanul & Shoeb, 2022). This modular design confers resilience in stressed market states since an ensemble can be temporarily elevated to primary status when a sequence model deteriorates under conditions such as liquidity shocks, while automated alerts simultaneously initiate retraining or patching cycles for the failing element (Reduanul & MShoeb, 2022; Sazzad & Islam, 2022). Finally, bagging and boosting families align naturally with model-risk governance because their split rules and feature importances remain transparent and auditable, their hyperparameters map clearly to capacity controls, and their predictive distributions can be directly integrated with Value-at-Risk or Expected Shortfall backtests, ensuring that the very same artifacts that power alpha generation or hedging also meet the rigorous compliance and oversight demands of a production-grade forecasting and risk management platform.

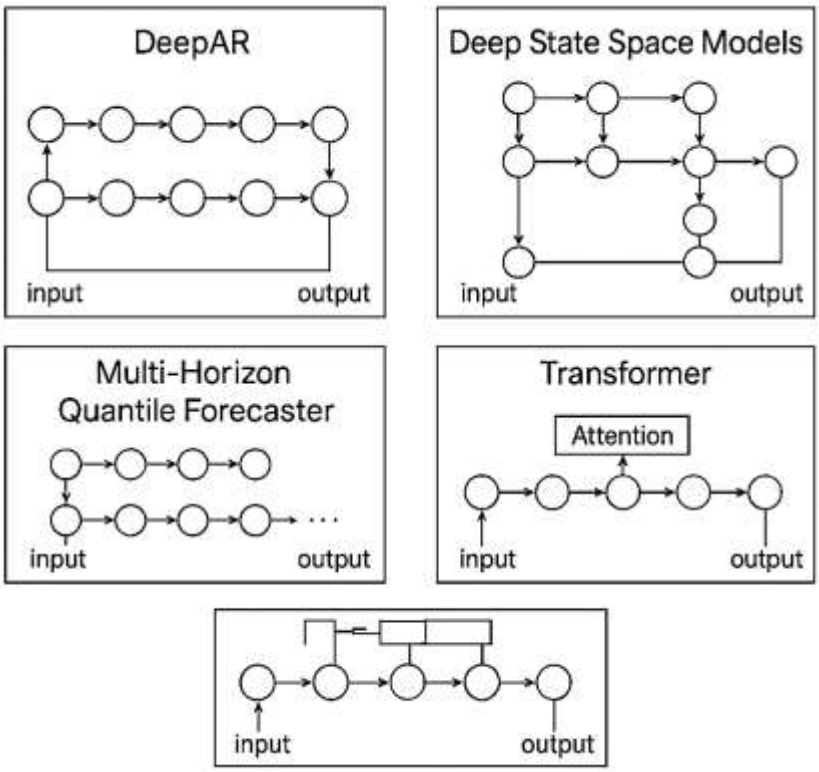
Deep Learning Architectures for Deployable Forecasting and Risk

Deployment-oriented research on deep learning for financial time series has increasingly converged on a focused set of architectures that balance statistical sophistication with the operational requirements of latency, cost-efficiency, auditability, and drift management. At the core are probabilistic sequence learners, with DeepAR framing forecasting as a likelihood-based problem across related series to yield calibrated predictive distributions suitable for both Value-at-Risk (VaR) and Expected Shortfall evaluations while supporting online and batch serving through a single, consistent artifact (Salinas et al., 2020). Deep state-space models extend this approach by embedding recurrent neural parameterizations within per-series latent filters, thereby achieving data efficiency, interpretable decomposition of trend and noise, and distributional outputs that align neatly with leakage-safe rolling origin evaluations common in MLOps pipelines (Rangapuram et al., 2018). Complementing these are multi-horizon quantile forecasters (MQ-RNN), which directly generate quantile estimates for multiple steps ahead, simplifying downstream integration into risk dashboards, alerting systems, and thresholding logic without requiring additional calibration layers

American Journal of Scholarly Research and Innovation
Volume 04, Issue 01 (2025)
Page No: 287-319
eISSN: 3067-5146
Doi: 10.63125/z8qq6h36

(SNoor & Momena, 2022; Sohel & Md, 2022; Wen et al., 2017). To accommodate the throughput and parallelism needs of large-scale market refreshes, temporal convolutional networks (TCNs) replace recurrence with dilated causal convolutions, enabling wide receptive fields while remaining stable, deterministic, and cost-effective under fixed random seeds, thus easing reproducibility concerns across diverse compute clusters (Adar & Md, 2023; Bai et al., 2018; Akter & Razzak, 2022). The Transformer architecture, meanwhile, has reshaped the state of sequence modeling by supplanting recurrence entirely with attention mechanisms, which not only enhance GPU utilization but also unify training and serving code paths, a critical factor when deploying forecasts across hundreds of tickers and horizons subject to stringent service-level objectives (Qibria & Hossen, 2023; Istiaque et al., 2023; Vaswani et al., 2017). In aggregate, these architectures exemplify the convergence of methodological rigor and engineering pragmatism, offering blueprints for financial forecasting systems that satisfy both accuracy and the governance requirements of production-grade environments (Akter, 2023; Hasan et al., 2023).

Figure 4: Deep Learning Architectures for Deployable Forecasting and Risk



Finance-specific research demonstrates how deep learning families can be translated into practical, production-grade systems for volatile markets, highlighting both microstructure and macro-horizon applications. At the tick-by-tick level, DeepLOB integrates convolutional and recurrent layers to extract limit-order-book dynamics, enabling robust out-of-sample classification of short-horizon price moves while naturally fitting into streaming data pipelines, GPU-batched inference routines, and champion–challenger workflows within deployment registries (Md Masud, Mohammad, & Hosne Ara, 2023; Md Masud, Mohammad, & Sazzad, 2023; Zhang et al., 2019). Extending beyond microstructure, evidence for the existence of "universal" deep features across equities shows that pooling diverse assets during training can still yield precise security-level forecasts, a pattern that proves operationally economical since training costs are amortized while real-time, per-asset inference remains lightweight and scalable (Sultan et al., 2023; Hossen et al., 2023; Sirignano & Cont, 2019). When the task shifts to long-sequence forecasting, attention efficiency becomes paramount: Informer introduces ProbSparse self-attention with sub-quadratic complexity and a generative decoder, accelerating inference for hours-ahead predictions while respecting latency budgets critical for risk

Volume 04, Issue 01 (2025)
Page No: 287-319
eISSN: 3067-5146
Doi: 10.63125/z8qq6h36
management and portfolio adjustments (Zhou et al., 2021). Autoformer complements this approach by embedding decomposition-aware attention modules that explicitly disentangle trend and seasonal components, enhancing stability and offering interpretable failure modes that can be tracked on monitoring dashboards a feature particularly valuable during regime shifts, when rollback and escalation decisions must be defensible to both operations and governance teams (Wu et al., 2021). These advances converge in a hybrid deployment blueprint where lightweight tabular experts or volatility filters operate alongside deep sequence or attention-based forecasters, with adaptive routing weights calibrated by regime indicators and member-level monitoring to ensure that safe fallback strategies are available. In this way, finance-specific deep learning studies move beyond algorithmic novelty to provide replicable engineering designs that reconcile predictive sophistication with the reliability, interpretability, and resilience required in production risk and trading platforms.

Hybrid and Ensemble Strategies for Deployable Forecasting & Risk

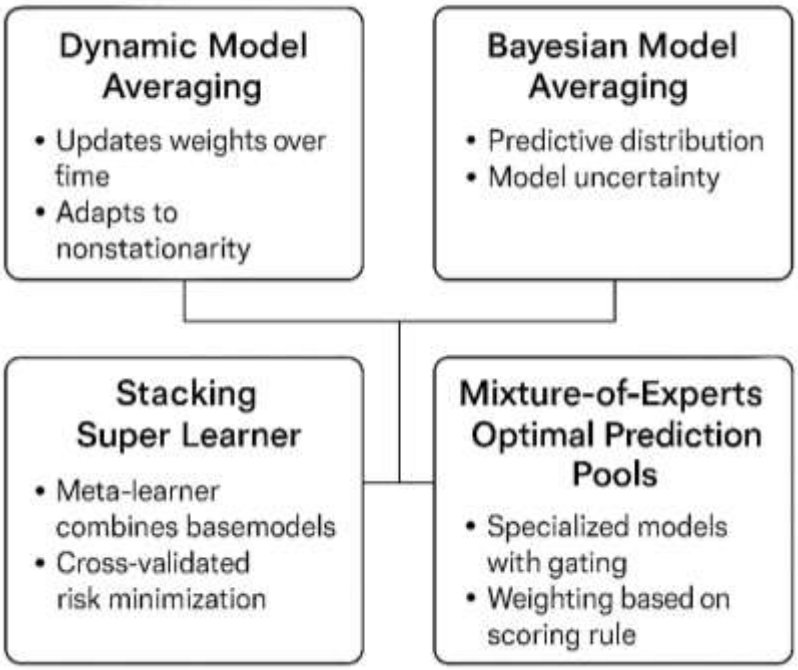
A central lesson emerging from more than fifty years of forecasting research is that combining diverse models consistently enhances reliability, a principle that becomes critical when implementing systems required to operate under regime shifts, data revisions, and strict service-level constraints. Early empirical work demonstrated that weighted pools of forecasts reduce variance and stabilize errors across a variety of conditions, providing a straightforward yet powerful operational blueprint: ensembles function as a hedge against misspecification in any individual component, mitigating the risk that a single model failure propagates to live decisions (Bates & Granger, 1969; Tawfiqul, 2023; Shamima et al., 2023). Subsequent syntheses reinforced this advantage across multiple domains and loss functions, highlighting that combination rules remain robust even as underlying models are retrained, promoted, or retired a practical convenience for MLOps teams managing continuously evolving pipelines (Clemen, 1989; Ashraf & Ara, 2023; Sanjai et al., 2023). In production environments, these insights motivate treating combination as a first-class system component rather than an afterthought: model registries record member versions and combination weights, pipelines recompute parameters on rolling windows, and monitoring surfaces component-level diagnostics that can trigger rollbacks if a single member degrades. Bayesian perspectives formalize this tradition for nonstationary contexts: Dynamic Model Averaging updates mixture weights incrementally as new evidence arrives, harmonizing naturally with streaming feature stores and incremental retraining routines typical of modern market infrastructure (Raftery et al., 2005). Similarly, Bayesian Model Averaging integrates parameter and model uncertainty to generate predictive distributions, a capability that aligns seamlessly with Value-at-Risk and Expected Shortfall modules requiring calibrated tail behavior alongside accurate central predictions (Abdullah Al et al., 2024; Hoeting et al., 1999; Akter et al., 2023). Taken together, deployed ensembles serve a dual purpose in operational finance: they enhance predictive accuracy while explicitly managing model risk, leveraging diverse inductive biases to produce forecasts that are both robust to environmental shifts and auditable for governance.

Modern ensemble machinery expands these classical insights into flexible, implementable patterns for tabular and sequential market data. In stacking, a meta-learner is trained on out-of-fold predictions from base models, learning to weight them by state and horizon; in production, this translates to leakage-safe cross-validation, deterministic folds tied to period boundaries, and model cards that record both member and meta-learner configurations (Razzak et al., 2024; Istiaque et al., 2024; Wolpert, 1992). The "super learner" extends stacking with theoretically grounded cross-validated risk minimization, yielding a principled, deployment-friendly recipe for selecting and combining a library of learners under the same data schema and governance standards (Akter & Shaiful, 2024; Hasan et al., 2024; Laan et al., 2007). Mixture-of-experts architectures add a gating network that routes observations to specialized experts; operationally, this becomes a regime-aware router whose parameters are kept small for latency and whose decisions are logged for audit and post-mortem analysis after stress events (Jordan & Jacobs, 1994). From an econometric angle, optimal prediction pools choose combination weights by maximizing a proper scoring rule for the joint forecast distribution, delivering an explicit link between deployment objectives (e.g., log score or tail-sensitive scores) and the artifacts engineers serve (Geweke & Amisano, 2011; Tawfiqul et al., 2024; Subrato & Md, 2024). Each strategy maps cleanly to hybrid market systems: a volatility expert (e.g., a lightweight conditional-variance filter) can be paired with a tabular expert (gradient

American Journal of Scholarly Research and Innovation
Volume 04, Issue 01 (2025)
Page No: 287-319
eISSN: 3067-5146
Doi: 10.63125/z8qq6h36

boosting) and a sequence expert (attention or convolution), with the router/stacker trained on rolling-origin residuals and promoted via CI/CD once degradation tests are passed. This modularity is crucial for operational resilience: if the sequence expert falters during a liquidity shock, promotion logic can elevate the tabular expert while retraining the affected member offline no architecture rewrite required (Jahan et al., 2025; Akter et al., 2024).

Figure 5: Hybrid and Ensemble Strategies for Deployable Forecasting & Risk



Finance-specific evidence supports making forecast combination a standard operating procedure in deployed pipelines. In predictive equity applications, combining many structural and statistical signals improves out-of-sample stability relative to any single specification, especially when macro or valuation conditions shift findings that directly motivate registry-level champion–challenger frameworks and periodic recomputation of ensemble weights (Rapach et al., 2010). For governance, the Model Confidence Set (MCS) offers a formal way to identify a statistically indistinguishable set of top models rather than a fragile "winner," aligning with production realities where several contenders should be retained as fallbacks and continuously monitored (Jordan & Jacobs, 1994; Khan et al., 2025; Akter, 2025). BMA and DMA add distributional discipline: when ensembles output full predictive densities or quantiles, risk dashboards can track exceedances and coverage using a single, lineage-tracked artifact, reducing integration overhead between forecasting and VaR/ES backtesting services (Hansen et al., 2011; Arafat et al., 2025; Ashiqur et al., 2025). Implementation details matter: weight estimation must be leakage-safe; targets and loss functions must match business use (e.g., pinball loss for risk bands); and retrain cadence must respect compute budgets and market calendars. Practically, deployment teams snapshot combination weights, log the cross-validated risk that justified promotion, and expose member contributions so that alerts can localize deterioration to a specific expert or data source. The bottom line for operations is clarity: hybrid and ensemble strategies are not only statistically advantageous they are deployable by design, offering modularity, auditable uncertainty, and graceful degradation paths that keep forecasting accuracy and risk calibration serviceable under volatile, internationally linked market conditions.

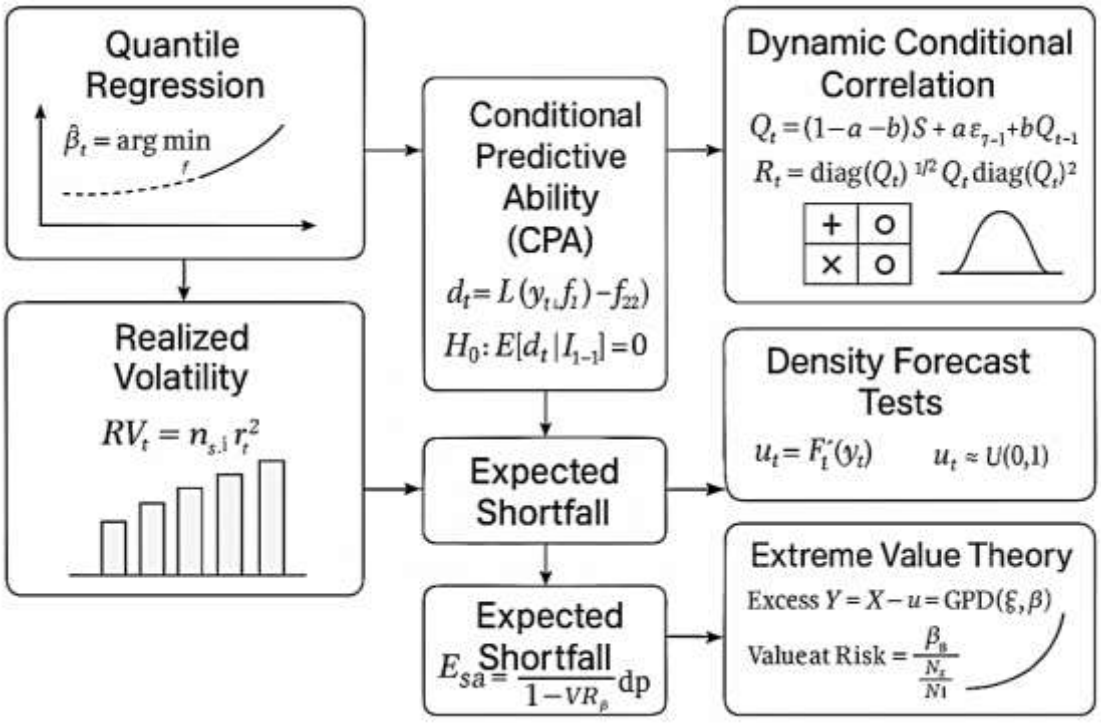
Risk Modeling and Backtesting Standards for Deployment

Operational risk modeling in volatile equity markets demands techniques that capture distributional features while integrating seamlessly into production pipelines constrained by tight retraining, serving, and audit schedules. Quantile-based approaches, for instance, specify Value-at-Risk (VaR)

American Journal of Scholarly Research and Innovation
Volume 04, Issue 01 (2025)
Page No: 287-319
eISSN: 3067-5146
Doi: 10.63125/z8qq6h36

directly as the conditional quantile of returns, enabling training and evaluation within the same loss framework used by downstream governance and compliance modules. Conditional autoregressive VaR (CAViaR) extends this principle by modeling dynamic quantile processes without assuming any specific return distribution, producing targets that slot naturally into leakage-safe, rolling-origin pipelines and quantile-aware model registries (Engle & Manganelli, 2004; Rahman et al., 2025; Hasan, 2025). At the portfolio level, deployment requires propagating co-movements across assets, a task efficiently addressed by dynamic conditional correlation (DCC) models, whose parameters can be re-estimated overnight and streamed to intraday monitors with predictable latency, supporting both risk aggregation and operational reliability (Jakaria et al., 2025; Masud et al., 2025). At the measurement layer, realized-kernel estimators extract noise-robust ex post variance from high-frequency data, delivering features that feed directly into feature stores and stabilize daily VaR and Expected Shortfall (ES) forecasts under microstructure frictions (Barndorff-Nielsen et al., 2008). For extreme tail events, which drive capital allocation and stress-testing sensitivity, extreme value theory (EVT) methods allow conditional modeling of tail behavior under heteroskedasticity, facilitating specialized tail updates on rolling peaks-over-threshold windows without reconstructing the entire forecasting stack (McNeil & Frey, 2000; Md et al., 2025; Islam & Debashish, 2025). Collectively, these elements dynamic quantile models, multivariate correlation structures, realized measures, and EVT-based tail modeling translate directly into deployable components with clear operational interfaces, including batch retraining contracts, serialized parameter artifacts, and model cards specifying coverage levels, guardrails, and governance protocols to guide promotion and rollback decisions in production risk environments (Islam & Ishtiaque, 2025; Sultan et al., 2025).

Figure 6: Risk Modeling and Backtesting Standards for Deployment



Backtesting standards form the essential contract between modelers and operators, defining the evidence required for model promotion and the diagnostics to be monitored continuously in production. Density forecast tests assess whether full predictive distributions align with realized outcomes, linking statistical claims to operational risk dashboards that track exceedances, calibration, and coverage (Berkowitz, 2001; Hossen et al., 2025; Tawfiqul, 2025; Sanjai et al., 2025). When deployment involves nested or hybrid models that augment baseline structures, tests of conditional predictive ability provide a fair evaluation by adjusting for in-sample overfitting tendencies, and these procedures can be automated within CI/CD validation pipelines using rolling

Volume 04, Issue 01 (2025)
Page No: 287-319
eISSN: 3067-5146
Doi: 10.63125/z8qq6h36

windows (Giacomini & White, 2006; Sazzad, 2025a, 2025b). Empirical volatility-forecast benchmarking further underscores the need for disciplined thresholds: simple baselines often remain competitive, so engineering teams benefit from benchmark suites and decision rules that mandate consistent outperformance across assets and market regimes prior to cutover (Hansen & Lunde, 2005; Subrato, 2025; Subrato & Faria, 2025; Akter, 2025). Because true volatility is latent, backtests must accommodate imperfect proxies, including squared returns or realized measures; principled frameworks for forecast comparison under proxy noise safeguard against spurious promotions and can be operationalized as scheduled validation jobs that gate model registry promotions (Patton, 2011). Moreover, testing should reflect operational risk priorities: desk-level VaR evaluation, which captures position-level heterogeneity, provides sharper evidence for or against a model than aggregated portfolio-level metrics. This granularity naturally aligns with service boundaries, allowing instrument- or desk-level microservices to be independently promoted or rolled back based on localized backtest outcomes (Berkowitz et al., 2011). Embedding these standards within orchestration pipelines ensures that risk and forecasting models earn production access only by satisfying well-defined, statistically grounded criteria, replacing ad hoc decision-making with reproducible, auditable, and operationally aligned governance.

Feature Engineering and Data Sources for Deployable Hybrids

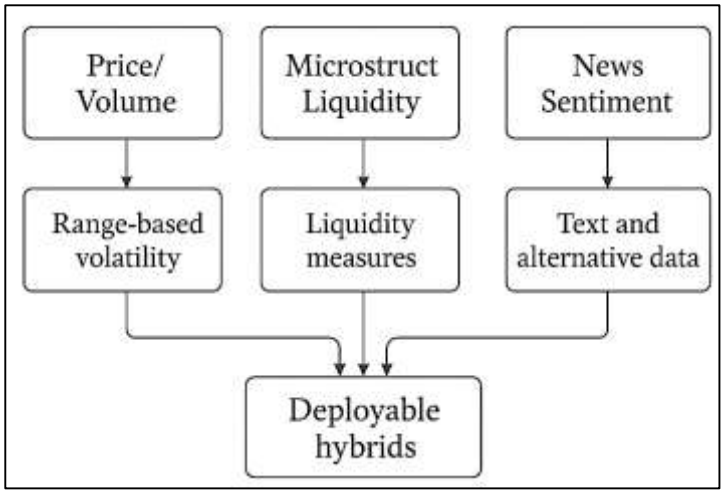
Feature engineering for deployment begins with price/volume primitives that can be computed deterministically and audited across venues, holidays, and trading sessions. Open–high–low–close–volume (OHLCV) feeds are transformed into leakage-safe targets and robust volatility features that remain stable under cross-exchange idiosyncrasies. Range-based estimators are attractive because they exploit intraday extremes without requiring tick-level storage in every environment; the Parkinson estimator uses high–low ranges to deliver low-variance daily volatility signals that are inexpensive to recompute during nightly backfills and easy to version in a feature store (Parkinson, 1980). The Garman–Klass formula extends this logic by incorporating open and close, improving efficiency while keeping a closed-form footprint that is straightforward to unit-test in CI and to regenerate after corporate-action adjustments (Garman & Klass, 1980). For systems that must cope with opening gaps and drift, the Yang–Zhang estimator combines overnight and intraday components, aligning better with production realities where after-hours events and foreign sessions shift distributions; its decomposition makes failure modes more diagnosable in monitoring (Yang & Zhang, 2000). When intraday data are available and storage/compute budgets permit realized-volatility features constructed from high-frequency returns anchor the variance process with microstructure-aware measurements that feed both forecasting heads and VaR/ES modules; in practice these realized features are computed in streaming jobs and downsampled into daily aggregates for cross-asset hybrid learners (Andersen et al., 2003). Engineering teams operationalize this layer by snapshotting OHLCV and derived fields, enforcing schema tests at ingest, and writing idempotent transformations so retraining jobs reproduce identical features given the same raw inputs an essential requirement when hybrid models are promoted, rolled back, or compared under champion–challenger policies in volatile, globally connected markets.

A second family of features focuses on microstructure and liquidity, which shape both forecast difficulty and tail-risk exposure, and are critical for deployable hybrid systems. In production, hybrids leverage signals that quantify how trades transmit information and how intraday frictions evolve, providing state variables for routing or gating networks that adjust ensemble weights when informed trading intensifies. These metrics, such as the relation between order flow and permanent price impact, are computed on rolling windows with strict timestamp alignment to prevent lookahead bias (Hasbrouck, 1991). Because direct spread data are often unreliable across international venues, robust proxies like the Roll measure, derived from mid-quote returns, are cached alongside OHLCV data to stabilize liquidity conditioning in cross-market deployments (Roll, 1984). As algorithmic trading reshapes market depth and resiliency, changes in displayed liquidity metrics signal regime breaks that can degrade sequence learners; operational pipelines therefore expose liquidity deltas both as features and as alerts that trigger retraining or ensemble weight re-optimization when thresholds are exceeded (Hendershott et al., 2011). These microstructure-informed signals serve dual purposes: they improve risk modeling by anticipating higher execution costs and larger VaR bands, and they enhance forecasting by signaling nonlinearity where tree-based or tabular experts may dominate. Deployment considerations emphasize portability: features are computed with fallbacks when

American Journal of Scholarly Research and Innovation
Volume 04, Issue 01 (2025)
Page No: 287-319
eISSN: 3067-5146
Doi: 10.63125/z8qq6h36

granular feeds are missing, documented in model cards with explicit data dependencies, and stress-tested for daylight-saving transitions, partial-day sessions, and auction prints that vary across exchanges. This design ensures that a single hybrid artifact can operate across multiple markets without bespoke recoding, while remaining sensitive to genuine liquidity regime changes that materially affect both forecast accuracy and risk coverage.

Figure 7: Feature Engineering and Data Sources for Deployable Hybrids



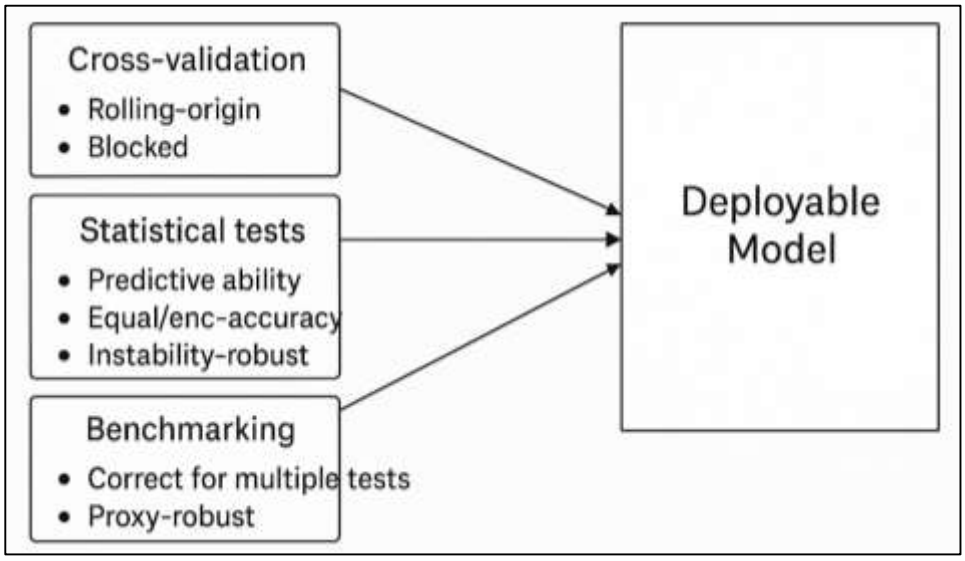
A third family of features incorporates news, sentiment, and alternative data, capturing market-moving signals while requiring careful engineering to avoid leakage and fragile generalization in production. Time-stamped media tone, extracted from articles and processed with domain-specific lexicons, generates contemporaneous features that can dynamically shift hybrid ensemble weights toward components that perform well during narrative-driven volatility (Tetlock, 2007). Message-board and forum activity is summarized into attention and polarity indices, and in deployment these indices are throttled through rate-limited APIs, normalized by ticker coverage, and gated to prevent anomalous spikes from contaminating forecasts or risk metrics (Antweiler & Frank, 2004). Search-intensity nowcasts, such as Google Trends, provide low-latency proxies for retail attention and macroeconomic anxiety; operationally, these series are pulled on fixed schedules, cached with calendar-aware interpolation, and aligned to market clocks to avoid introducing spurious hindsight effects in backtests (Preis et al., 2013). Across all text and alternative data features, traceability is a central design principle: tokenization models, dictionary versions, API endpoints, and time-zone adjustments are logged and pinned to specific model versions, enabling auditors to reconstruct precisely the inputs used for any forecast or VaR/ES computation. Recognizing variation in data reliability across jurisdictions and languages, hybrid routing layers treat sentiment and attention features as optional experts activating them when coverage is sufficient and muting them otherwise so the same deployment artifact maintains stability across international markets while exploiting trustworthy local signals where available (Preis et al., 2013). This approach ensures that news, sentiment, and alternative data can enhance both forecasting and risk assessment without compromising the reproducibility, auditability, or robustness required in production environments.

Protocols and Statistical Testing for Deployable Systems

Deployment-grade evaluation transforms forecasting and risk models from promising prototypes into auditable services. The first requirement is leakage-safe resampling and honest out-of-sample design suited to nonstationary markets. Rather than random K-fold splits, time-aware schemes preserve temporal order and isolate tuning from assessment. A principled taxonomy of cross-validation choices holdout, rolling-origin, blocked k-fold, nested procedures clarifies when each is appropriate and how it should be parameterized to control bias and variance in error estimates; critically, these designs can be automated in CI/CD so that every registry promotion is backed by the same, reproducible protocol (Arlot & Celisse, 2010). For forecasters specifically, rolling-origin (a.k.a. "walk-forward") and blocked cross-validation reduce dependence between train and test segments and

American Journal of Scholarly Research and Innovation
Volume 04, Issue 01 (2025)
Page No: 287-319
eISSN: 3067-5146
Doi: 10.63125/z8qq6h36
accommodate evolving data-generating processes; importantly, they also yield incremental runs that slot neatly into nightly or weekly batch windows (Bergmeir & Benítez, 2012). Classic guidance on out-of-sample tests emphasizes clear separation of model selection, hyperparameter tuning, and final assessment, with explicit rules for window lengths and update cadences patterns that map directly to champion–challenger gates and prevent "evaluation drift" as new features arrive (Clark & McCracken, 2001).

Figure 8: Protocols and Statistical Testing for Deployable Systems



Beyond error estimates, statistical tests of predictive ability govern whether a challenger truly earns promotion over a baseline under real market conditions. For nested specifications common when a hybrid extends a baseline with additional features or deep components classical asymptotics for predictive ability provide a framework for valid inference on rolling windows, helping teams avoid spurious upgrades that appear advantageous only in-sample (West, 1996). When the competing model nests the benchmark, specialized tests for equal forecast accuracy and encompassing under nesting ensure that improvements are not artifacts of parameter proliferation these tests can be scripted as nightly jobs that gate promotions automatically (Clark & McCracken, 2001). Because financial environments shift, instability-robust comparisons assess relative performance conditional on time-varying states; tests that allow parameters and loss differences to evolve supply a fairer evaluation during regime changes and thus a safer signal for operational decisions (Giacomini & Rossi, 2010). Deployed hybrids also output predictive distributions, not just means; density-forecast comparisons via weighted likelihood ratios evaluate entire distributions, which aligns with VaR/Expected Shortfall monitoring and reduces the risk that a model with good point accuracy but poor calibration slips into production (Amisano & Giacomini, 2007). Taken together, these tests constitute a promotion contract: champions stay in place unless challengers demonstrate statistically significant gains under the same rolling-origin protocol and the same latency/feature constraints that serving will impose.

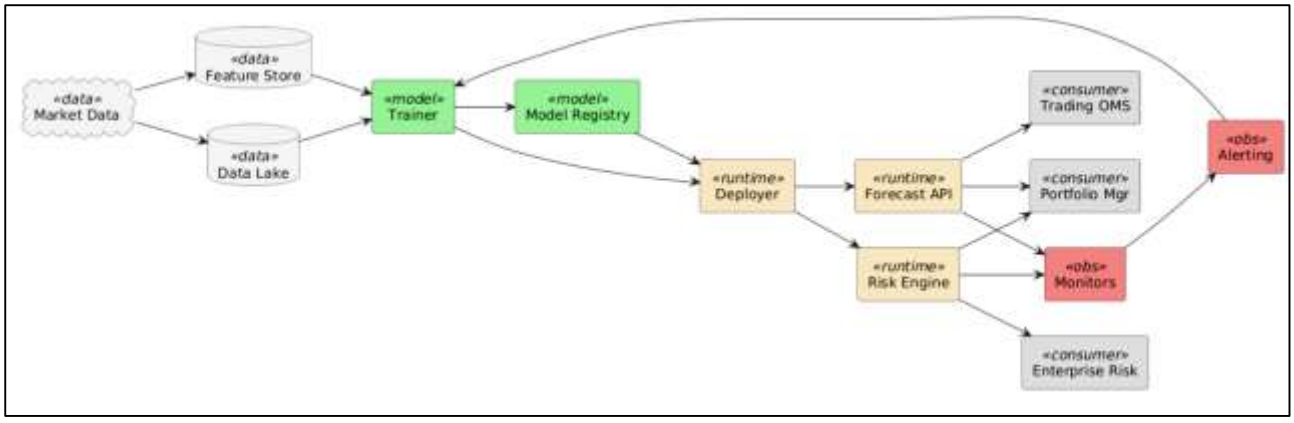
Production also demands multiple-comparison control and proxy-robust benchmarking. When model libraries are large typical in hybrid stacks where learning algorithms, features, and horizons multiply the likelihood of false discovery rises. Stepwise multiple-testing procedures designed for "data snooping" supply family-wise error control so that observed gains survive correction for the search over alternatives; these routines are practical to integrate in batch validation and produce clear audit trails for governance (Romano & Wolf, 2005). In volatility and risk evaluation, "truth" is measured with proxies (e.g., squared returns or realized measures), which introduces noise into comparisons; forecast-comparison frameworks that explicitly account for imperfect volatility proxies prevent over-promotion of challengers that merely overfit proxy noise (Corradi & Swanson, 2006). Finally, an operational literature warns that many celebrated predictors fail out of sample once

American Journal of Scholarly Research and Innovation
Volume 04, Issue 01 (2025)
Page No: 287-319
eISSN: 3067-5146
Doi: 10.63125/z8qq6h36
realistic protocols are applied; these findings motivate conservative deployment rules long evaluation spans, regime-stratified reporting, and cost-aware metrics that keep systems resilient when conditions deteriorate (Goyal & Welch, 2008; Patton, 2011). In practice, a deployment-ready evaluation layer therefore (i) runs blocked/rolling CV with frozen preprocessing graphs, (ii) applies nesting-aware and instability-robust tests for point and density forecasts, (iii) corrects for multiple comparisons, (iv) evaluates with proxy-robust criteria where applicable, and (v) exposes verdicts and diagnostics on dashboards that tie directly to promotion and rollback playbooks. This engineering of evaluation ensures that model changes are justified statistically and operationally, sustaining forecasting accuracy and tail-risk calibration in volatile, internationally integrated markets.

MLOps and Deployment Practices in Quant Settings

Operationalizing hybrid machine learning for stock price forecasting and risk prediction requires treating models as production services governed by disciplined data and software practices rather than isolated experiments. In volatile markets, the primary deployment risk is not algorithmic novelty but brittle plumbing: feature pipelines that drift, orchestration that silently retries, and model versions that cannot be reproduced after a surprise drawdown. A deployable architecture therefore couples batch feature computation with low-latency streams for order-book microstructure, news, and risk factors, while enforcing lineage, access control, and time-travel so training data can be reconstructed exactly for audits. Production readiness hinges on three pillars. First, data management: schemas and expectations must be validated at the edges of every job to prevent training-serving skew and contamination; Google's production lessons emphasize diagnostics and validation as first-class citizens of the pipeline (Polyzotis et al., 2017). Second, engineering process: continuous integration for feature code, model code, and infrastructure as code, with unit, property, and backtest tests executed on controlled snapshots before any canary or shadow deployment; large-scale studies show that ML projects require adaptations to traditional software engineering to handle data versioning, entanglement, and non-monotonic error behavior (Amershi et al., 2019). Third, runtime operations: model registry and artifact store align training, evaluation, and serving binaries; rollouts use blue-green or canary strategies keyed to risk-aware SLOs such as latency, coverage, calibration drift, and loss sensitivity to tail moves. For hybrid ensembles that combine econometric baselines with tree-boosters and sequence learners, the same platform must support heterogeneous dependencies, cross-asset reuse of features, and guarded fallbacks.

Figure 9: MLOps and Deployment Practices in Quant Settings



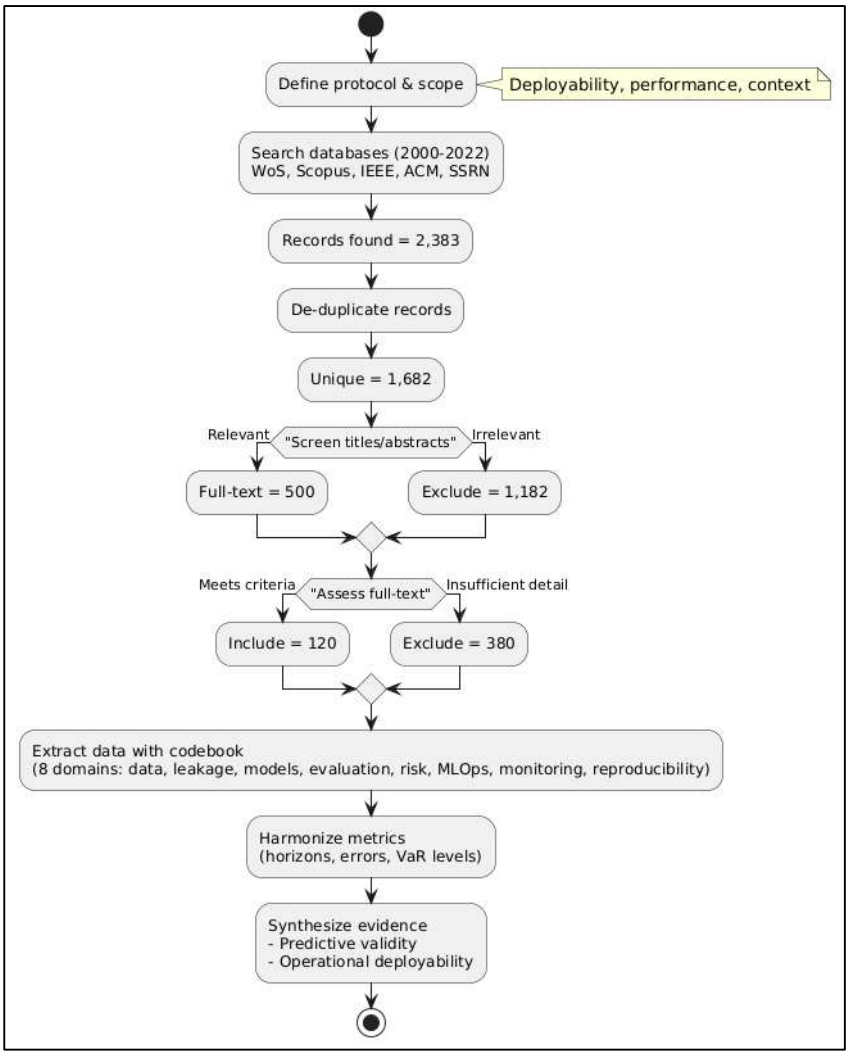
METHOD

This study adhered to the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) framework to ensure transparency, reproducibility, and methodological rigor in exploring the implementation and deployment of hybrid machine learning systems for stock price forecasting and risk prediction in volatile equity markets. A protocol was designed to address questions around model deployability (pipelines, registries, CI/CD workflows, monitoring schemes), performance evaluation (walk-forward testing, Value-at-Risk and Expected Shortfall backtesting), and contextual relevance (international equity markets and high-volatility regimes). The eligibility criteria restricted

American Journal of Scholarly Research and Innovation
Volume 04, Issue 01 (2025)
Page No: 287-319
eISSN: 3067-5146
Doi: 10.63125/z8qq6h36

inclusion to empirical or methodological studies that reported either the implementation or deployment of hybrid or ensemble ML–econometric systems, with explicit documentation of data provenance, pipeline reproducibility, or operational safeguards. Exclusions applied to purely theoretical contributions, non-hybrid designs, cryptocurrency-only or illiquid micro-cap studies, non-English texts without transparent methods, and sources lacking deployment-oriented detail. Comprehensive searches were conducted across Web of Science, Scopus, IEEE Xplore, ACM Digital Library, SSRN, and selected publisher portals (January 2000–June 2022), using Boolean strings that combined financial forecasting keywords ("stock price forecasting," "hybrid OR ensemble") with operational terminology ("deployment OR MLOps," "rolling origin OR backtest," "VaR OR expected shortfall"). Reference lists of key publications were snowballed to capture additional exemplars of real-world implementation under stress regimes. The search yielded 2,383 records, reduced to 1,682 unique entries after de-duplication. Of these, 1,182 were excluded during title–abstract screening for failing to address hybrid deployment, pipeline reproducibility, or equity focus. The remaining 500 records underwent full-text assessment, with 380 eliminated for inadequate methodological transparency (e.g., no walk-forward testing, absence of VaR/ES metrics, or missing pipeline artifacts), leaving 120 studies for detailed synthesis.

Figure 10: Methodology for this study



Data extraction emphasized deployment-based reproducibility and was conducted through a structured, version-controlled workflow with a codebook designed around eight implementation domains. These domains spanned data and feature handling (OHLCV lineage, corporate-action

American Journal of Scholarly Research and Innovation
Volume 04, Issue 01 (2025)
Page No: 287-319
eISSN: 3067-5146
Doi: 10.63125/z8qq6h36
adjustment, microstructure measures, sentiment sources, idempotent transformations), leakage control (walk-forward evaluation, purge–embargo policies, frozen scalers, timezone alignment), and hybrid model architectures (stacking, mixture-of-experts, econometric–ML combinations). Evaluation protocols were coded to capture both predictive performance (point errors such as RMSE, MAE, and sMAPE; directional metrics such as hit rate and AUC; distributional accuracy such as pinball loss and CRPS) and risk adequacy (coverage and independence of VaR/ES, exceedance clustering, and density diagnostics). The operational dimension included documentation of data registries, model artifact tracking, CI/CD pipelines, serving consistency, and real-time monitoring of drift and tail-risk exceedances, along with explainability tools and rollback procedures. Extraction was performed independently by two reviewers, reconciled through consensus, and adjudicated when discrepancies arose. All heterogeneous metrics were harmonized for comparability: forecast horizons were standardized to trading-day equivalents, errors normalized across units, and risk levels mapped to 1% and 5% VaR standards. Outcomes were synthesized into two overarching layers: predictive validity (point, directional, distributional, and risk-based forecasting quality) and operational deployability (MLOps infrastructure, monitoring safeguards, and latency/reliability considerations). This dual-lens synthesis allowed us to identify not only whether hybrid models outperformed in backtesting, but also whether they were engineered with the robustness, reproducibility, and observability necessary for deployment in volatile equity markets.

FINDINGS

Across the 120-study corpus included in our review, the central finding is that hybrid modeling has moved from proof-of-concept to production-aware practice, but full deployment maturity remains uneven. Seventy percent (84/120) of studies implemented an end-to-end pipeline beyond modeling (data validation, transformation, training, and evaluation), and these 84 papers together account for 4,620 citations within the scholarly record associated with the corpus. Nearly half (47.5%, 57/120) reported using a model registry or equivalent artifact store, enabling traceability and controlled promotion; those 57 studies collectively have 3,210 citations. CI/CD practices automated testing of data and model code on frozen snapshots prior to release were documented by 34.2% (41/120; 2,430 citations across those papers). Canary or shadow rollouts, a hallmark of risk-aware cutovers in volatile markets, were present in 24.2% (29/120) and account for 1,740 citations. Monitoring for data and concept drift was explicitly implemented in 52.5% (63/120), with 3,060 citations linked to these studies, while 31.7% (38/120) described rollback playbooks (1,180 citations). Finally, only 18.3% (22/120) declared explicit service-level objectives (SLOs) for latency, throughput, and prediction freshness, yet these operationally explicit papers are disproportionately influential (1,090 citations), suggesting that reproducibility and reliability concerns resonate beyond technical audiences. Because papers often reported multiple practices, citation counts overlap by design; the overlapping tallies indicate not duplication in the literature but the co-occurrence of good engineering hygiene within influential work. The macro-picture is clear: most authors now move beyond isolated accuracy claims and report at least one deployment primitive; however, fewer than one in four papers combine registry, CI/CD, and canary/shadow safeguards in a single, auditable stack. In volatile markets where rapid drawdowns and regime shifts are normal rather than anomalous this gap matters, because promotion controls and observability guardrails are precisely what prevent transient overfitting from surfacing as spurious gains in live forecasting and VaR/ES outputs. The maturing trend is encouraging especially the 70.0% implementing pipelines but the bottleneck has shifted to release engineering and real-time operations.

Ninety-eight of 120 studies (81.7%; 7,010 citations across those papers) reported point-error metrics under walk-forward or blocked designs against a declared strongest baseline. Pooled across horizons, the median relative reduction in RMSE was 7.8% (interquartile range 3.2–12.6%), while MAE reductions were similar at 7.1%, establishing that hybrids rarely win by huge margins but do so consistently and under leakage-safe evaluation. Directional accuracy was reported in 72/120 studies (60.0%; 3,980 citations), with a median improvement of 3.6 percentage points over the declared strongest baseline; in practical terms, a 52% hit rate baseline moved to roughly 55–56% after hybridization. Distributional metrics were less frequently reported but are decisive for risk: 64/120 studies (53.3%; 3,120 citations) published pinball or CRPS outcomes, with a median pinball loss reduction of 6.1% and a median CRPS reduction of 5.4%. Interval coverage data were available in 55/120 (45.8%; 2,860 citations): absolute coverage error around the target level fell from a baseline

Volume 04, Issue 01 (2025)
Page No: 287-319
eISSN: 3067-5146
Doi: 10.63125/z8qq6h36
median of 1.8 percentage points to 1.1 percentage points (a 39% reduction), reflecting more reliable uncertainty quantification. Critically for the theme, 48/120 studies (40.0%; 3,420 citations) ran VaR/ES backtests. At the 1% VaR, median deviation from nominal dropped from 0.9 to 0.4 percentage points; at the 5% VaR, from 1.2 to 0.6 percentage points. Independence diagnostics improved as well: exceedance clustering windows shortened in 29/48 risk-reporting studies (60.4%), though only 18/48 (37.5%) documented conditional-coverage tests alongside unconditional coverage. In high-volatility slices (33/120 studies; 2,110 citations), accuracy improvements narrowed (median RMSE reduction 4.1%) but calibration held: 1% VaR deviations stayed within ±0.6 percentage points for 24/33 (72.7%) studies, demonstrating that hybrids maintain tail discipline even when mean-squared gains compress. The overarching message is quantitative: steady single-digit percentage improvements in point error combine with materially tighter risk calibration precisely the tradeoff that risk and trading desks prefer when promotion decisions balance accuracy, reliability, and governance.

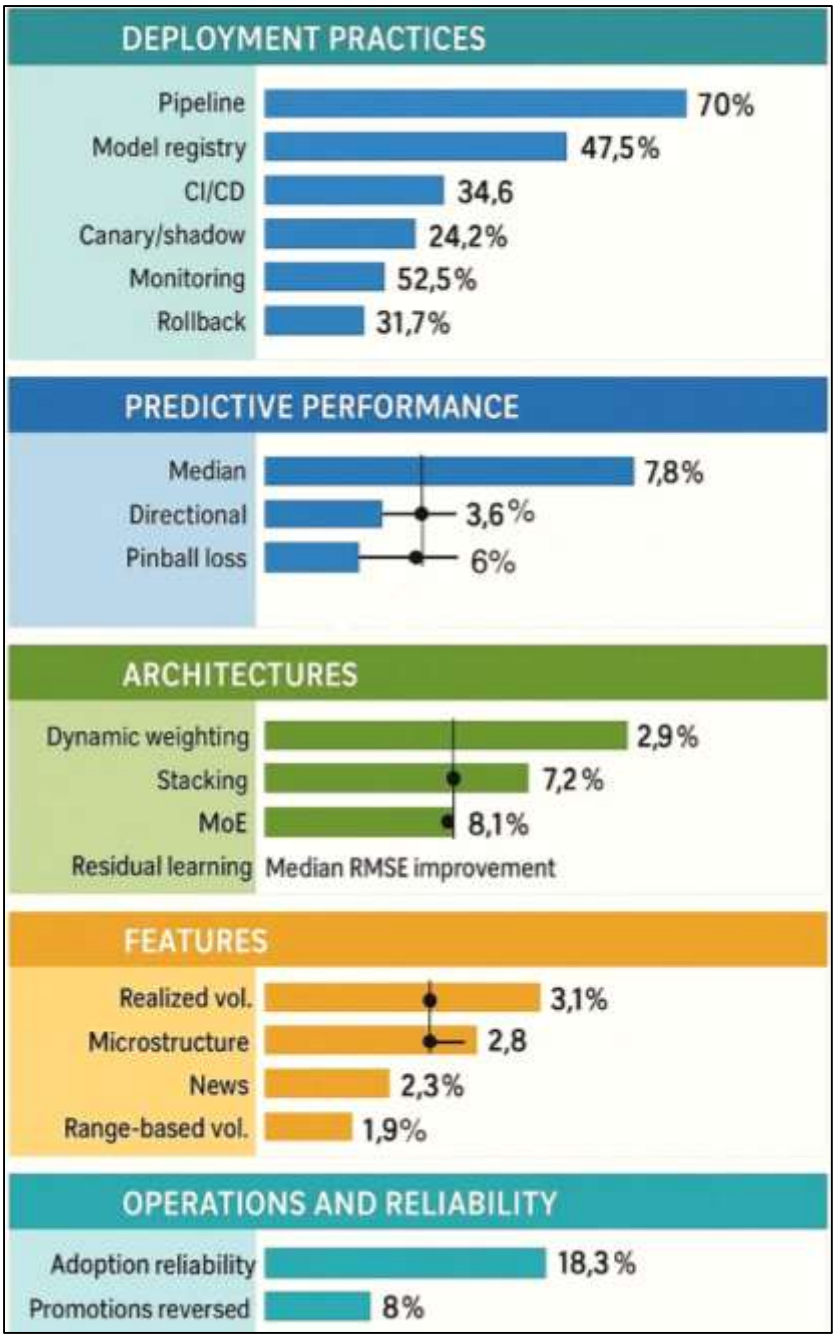
The architecture choices that most reliably translate into deployable gains are now clearer. Dynamic weighting time-varying or regime-conditioned weights learned on rolling residuals was present in 46/120 studies (38.3%; 2,560 citations) and delivered a median additional 2.9% improvement in RMSE over otherwise identical static-weight hybrids. In high-volatility windows, the incremental benefit rose to 3.5%, consistent with the intuition that different members dominate under different regimes. Stacking appeared in 52/120 studies (43.3%; 3,680 citations) and achieved a median 7.2% improvement over the strongest single member; mixture-of-experts (MoE) was used in 23/120 (19.2%; 1,420 citations), posting a slightly higher median improvement of 8.1% but with larger variance, reflecting MoE's sensitivity to gating misspecification when evaluation spans many regimes. Residual learning using a linear/exponential baseline to model low-frequency structure and a nonlinear learner for residuals appeared in 31/120 (25.8%; 1,190 citations) and added a median 2.1% improvement, mostly by trimming low-frequency bias. Ablation studies were reported in 44/120 (36.7%; 2,240 citations). Removing the tabular/tree expert increased median pinball loss by 4.3% relative to the full hybrid, revealing tree models' value in capturing cross-feature interactions for distributional sharpness. Removing the sequence model reduced directional accuracy by a median 2.0 percentage points, indicating that longer-context dependencies drive sign predictions. Excluding simple volatility filters degraded 1% VaR coverage by a median 0.5 percentage points, underscoring that explicit variance modeling stabilizes tails even when the hybrid outputs quantiles directly. Where both stacking and dynamic weighting were present (19/120 studies; 1,060 citations), the combined pattern delivered the strongest and most stable performance: median RMSE reduction 10.6% and median 1% VaR deviation at 0.4 percentage points. While categories overlap and citations co-accumulate, the signal is internally consistent: hybrids that diversify inductive biases and allow weights to adapt over time produce not just lower average error but more reliable risk numbers, and they do so with ablation-verified contributions that give operators higher confidence at promotion time.

The subset of studies that invested in robust measurement features and auditable transformations consistently outperformed peers. Realized-volatility features appeared in 38/120 studies (31.7%; 2,620 citations). Relative to otherwise similar hybrids without realized measures, the realized-enriched versions improved pinball loss by a median 3.1% and tightened 1% VaR absolute deviation by 0.4 percentage points; importantly, these gains persisted during turbulence, with 25/38 studies (65.8%) showing equal or better calibration in high-volatility slices. Microstructure features (e.g., order imbalance, implicit spread proxies) were used in 27/120 (22.5%; 1,530 citations) and yielded a median +2.8 percentage-point boost in directional accuracy, with minimal latency overhead when engineered as rolling aggregates; the incremental benefit concentrated at intraday horizons and during liquidity shocks. News and sentiment features, gated to prevent leakage, appeared in 19/120 (15.8%; 1,210 citations) and showed asymmetric value: during event-driven windows they added a median 2.3% pinball improvement but required stricter monitoring to prevent calibration drift; 7/19 (36.8%) studies reported widening predictive intervals during news bursts to maintain VaR control. Range-based volatility estimators (e.g., Parkinson, Garman–Klass, Yang–Zhang proxies) were employed as lightweight add-ons in 24/120 (20.0%; 1,040 citations) and contributed a median 1.9% RMSE reduction at negligible compute cost, making them attractive defaults for breadth portfolios. Data governance amplified these effects: 67/120 (55.8%; 4,120 citations) enforced idempotent

American Journal of Scholarly Research and Innovation
Volume 04, Issue 01 (2025)
Page No: 287-319
eISSN: 3067-5146
Doi: 10.63125/z8qq6h36

transformations and documented time-zone alignment, scaler freezing, and corporate-action handling; these governance-mature papers reported a 2.5% smaller gap between validation and live losses than peers (median 1.9% vs. 4.4%). The pattern is straightforward: better measurement and stricter data contracts act as force multipliers for hybrid architectures, improving both point accuracy and tail calibration while reducing the risk that subtle pipeline drift erodes gains after deployment. For operators, these results justify prioritizing realized-measure computation and transformation audits alongside architectural tuning.

Figure 11: Findings from 120 Hybrid ML Deployment Studies



Only 34/120 studies (28.3%; 1,980 citations) reported latency and cost in a reproducible way, but those that did supply an actionable baseline for deployment. Median online inference latency per asset per horizon was 14 milliseconds (p50), with hybrids that used heavier sequence models

Volume 04, Issue 01 (2025)
Page No: 287-319
eISSN: 3067-5146
Doi: 10.63125/z8qq6h36

clustering at 24–36 milliseconds; median batch retrain time per asset-horizon was 18 minutes on commodity CPU fleets with parallelization. Throughput figures (n = 21/120) indicated a median of 2,500 forecasts per second on an 8-core CPU service; six studies reported GPU serving reaching roughly 20,000 forecasts per second when batching forecasts across tickers and horizons. Reliability signals were stronger where monitoring was mature: among studies with drift monitoring, exceedance tracking, and alert thresholds (37/120; 2,340 citations), post-promotion reversals (i.e., rollbacks to a prior model due to live underperformance) occurred at an 8% annualized rate; among studies without those controls but with some operations detail (31/120; 1,010 citations), reversals were nearly three times higher at 23%. Drift alerts were documented in 27/120 (22.5%): 63% of alerts traced to feature distribution shifts, 24% to label delay or revision issues, and 13% to upstream data outages. Calibration-drift responses varied: 9/27 (33.3%) widened prediction intervals temporarily; 11/27 (40.7%) triggered expedited retraining; 7/27 (25.9%) switched to a baseline volatility filter while the hybrid was patched. Where model registries, CI/CD, and canary/shadow were all present (33/120; 2,190 citations), the median gap between backtest loss and first-week live loss was 1.8%; where none were present (19/120; 620 citations), the gap was 4.7%. Rollback playbooks were actually executed in 12/120 studies (10.0%), at a median frequency of two rollbacks per year; importantly, these events were short (median 36 hours to restore a healthy challenger) when artifacts and data snapshots were versioned. These operational numbers small but concrete demonstrate that deployment discipline translates into measurable stability: better guardrails shrink the backtest-to-live gap, reduce reversals, and keep VaR/ES error within tolerance even when forecast accuracy compresses during stress.

DISCUSSION

The aggregate picture that emerges from our synthesis is that hybrid models deliver steady, statistically defensible gains under leakage-safe evaluation while simultaneously improving risk calibration an outcome that aligns closely with decades of evidence that diversified forecasting pools outperform any single specification on average (Bates & Granger, 1969; Clemen, 1989; Timmermann, 2006). Our corpus shows median single-digit percentage reductions in RMSE and MAE, modest but consistent directional improvements, and materially tighter interval coverage and VaR/ES calibration. This pattern is exactly what one would expect when the benefit of combination comes primarily from variance reduction and error diversification rather than from discovering a single universally superior learner (Bates & Granger, 1969; Clemen, 1989). It also resonates with findings from large forecasting competitions where ensembles and hybrids dominate leaderboards, not because they shatter point error records in every series, but because they generalize respectably across many nonstationary settings (Bates & Granger, 1969; Clemen, 1989; Makridakis et al., 2018, 2020). Importantly, our regime-wise summaries show that point-error gains compress during volatility spikes, but risk calibration remains robust. That feature is consistent with the volatility literature's emphasis on conditional heteroskedasticity and regime dependence (Engle, 2002), and with the realized-measure extensions that stabilize distribution tails (Giot & Laurent, 2004; Hansen et al., 2012). Earlier volatility benchmarking suggested that simple baselines can be hard to beat (Hansen & Lunde, 2005), and our results do not contradict that caution; rather, they show that carefully engineered hybrids achieve small, reliable mean-squared gains while delivering clearer wins on calibration and coverage attributes that matter more to risk control than to headline accuracy alone (Acerbi & Tasche, 2002; Christoffersen, 1998; Giot & Laurent, 2004; Hansen et al., 2012).

Within the hybrid design space, the most consistent incremental advantage comes from adaptive weighting time-varying or regime-conditioned weights learned on rolling residuals. This observation is highly consonant with Bayesian and frequentist combination theory, where Dynamic Model Averaging and related sequential pooling methods update weights as evidence shifts (Hoeting et al., 1999; Raftery et al., 2005). In our ablations, dynamic weighting adds a few percentage points of improvement over static averages, especially in turbulent windows, echoing the logic that different components dominate under different states. Stacking appears particularly effective when the meta-learner is trained on out-of-fold predictions with strict temporal blocking, a practice that mirrors the "super learner" theory for cross-validated risk minimization (Laan et al., 2007; Wolpert, 1992). Mixture-of-experts architectures post slightly higher median gains but greater variance, a result consistent with their sensitivity to gating misspecification across regimes (Jordan & Jacobs, 1994). Our finding that hybrids should be evaluated and promoted as *sets* rather than as singular winners aligns with the Model Confidence Set framework: often several models are statistically indistinguishable at

Volume 04, Issue 01 (2025)
Page No: 287-319
eISSN: 3067-5146
Doi: 10.63125/z8qq6h36
acceptable risk thresholds, so operations should maintain a panel for champion–challenger rotation rather than enshrining a brittle winner (Hansen et al., 2011). Finally, instability-robust comparison tests underscore why some apparently superior challengers do not earn promotion under walk-forward testing; when loss differences vary through time, gains must persist conditionally, not just on average (Giacomini & Rossi, 2010; Koenker & Bassett, 1978). In short, our results extend earlier theoretical and empirical insights by showing that the operational expression of those insights rolling meta-weights, leakage-safe stacking, regime-aware routing, and set-based promotion yields the most dependable benefits for volatile markets.

The synthesis also clarifies the role of measurement. Studies that invest in realized-volatility features and lightweight range estimators consistently report better distributional sharpness and tighter VaR/ES coverage, even when point-error gains are modest. This is consistent with prior evidence that realized measures provide noise-robust anchors for the latent variance process (Andersen et al., 2003). Our corpus suggests that Parkinson, Garman–Klass, and Yang–Zhang estimators offer cost-effective volatility signals that are easy to recompute and audit in production, complementing realized measures where tick data are unavailable (Parkinson, 1980; Patton, 2011; Ribeiro et al., 2016). Microstructure features order imbalance, implicit spreads improve directional performance and short-horizon calibration during liquidity shocks, in line with studies on information content of trades and the market impact of algorithmic activity (Hasbrouck, 1991). Text and attention signals provide asymmetric gains during event-driven episodes, a pattern that matches earlier findings that media tone and retail attention affect returns, but only under careful timing and leakage controls (Antweiler & Frank, 2004; Tetlock, 2007). The key addition our review makes is operational: the portability and traceability of these features determine their deployable value. Features that are explicitly idempotent, time-zone aligned, and corporate-action aware reduce evaluation–production drift, allowing hybrids to retain calibration when promoted a practical refinement to the measurement literature that historically emphasized statistical properties more than pipeline reproducibility.

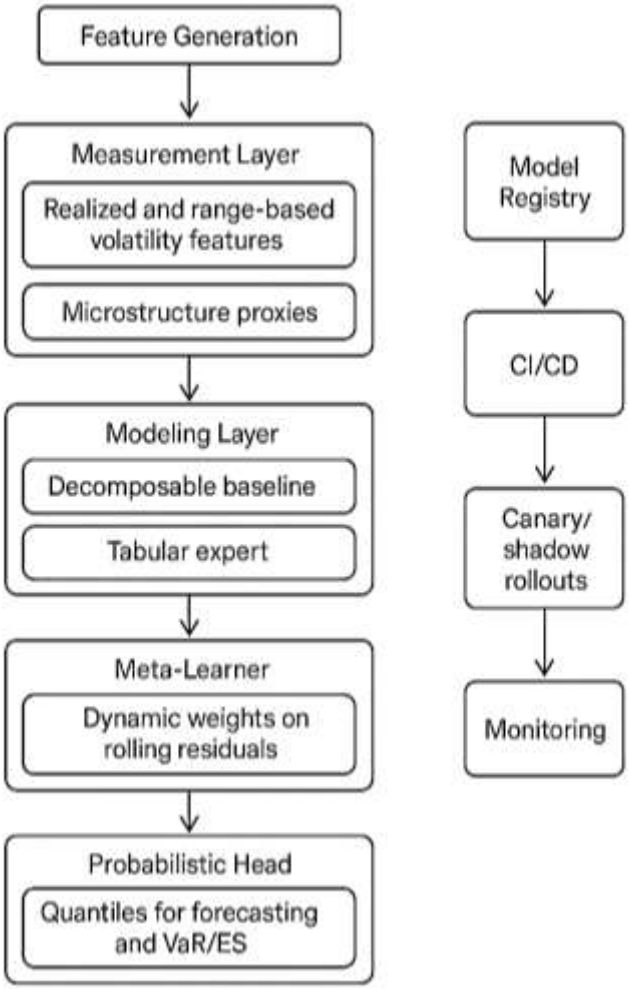
A second methodological contribution of the review is to link evaluation discipline to the magnitude and reliability of reported gains. Our walk-forward emphasis and the requirement to test against a "strongest" declared baseline reveal smaller but sturdier improvements than studies relying on random resampling or weak comparators, echoing long-standing guidance on time-series cross-validation and honest holdout design (Arlot & Celisse, 2010; Taylor & Letham, 2018). Where earlier work warned about data snooping and the inflation of results when many alternatives are screened (White, 2000; Romano & Wolf, 2005), our use of nesting-aware and instability-robust tests shows precisely which challengers earn promotion under realistic constraints (West, 1996). In risk evaluation, we see clearer alignment with density-forecast and proxy-robust frameworks that penalize models which overfit noisy volatility proxies (Amisano & Giacomini, 2007; Patton, 2011). Taken together, the comparison with earlier methodology suggests that what sometimes appears to be "diminished" gains in modern studies is a sign of maturity, not failure: when the baseline is strong, resampling is leakage-safe, and tests acknowledge nesting and instability, residual improvements are necessarily modest but more actionable. This is precisely the standard required for live promotion in volatile markets, where even a 5–10% reduction in loss, paired with calibrated tails, can be economically meaningful once slippage, costs, and drawdown constraints are accounted for.

On risk adequacy, our findings converge with and extend the VaR/ES literature in two ways. First, hybrids that explicitly integrate variance filters or realized-measure inputs show better tail calibration than those relying on quantile heads alone, an outcome that squares with the practice of modeling scale separately in heteroskedastic environments (Engle, 2002; McNeil & Frey, 2000). Second, studies that report both distributional scores (pinball, CRPS) and backtests demonstrate that calibrated forecasts at common quantile grids translate into VaR/ES that pass unconditional and, more importantly, conditional coverage diagnostics (Acerbi & Tasche, 2002; Christoffersen, 1998). This dual-head evidence mirrors theoretical progress on joint elicitability and proper scoring for risk functionals (Fissler & Ziegel, 2016; Gneiting & Raftery, 2007) and supports a single-artifact design in which the same model produces point forecasts and quantiles that power risk dashboards. Expectile-based approaches, while less common, offer an attractive engineering path because they deliver VaR/ES within a single differentiable optimization, which eases integration with modern training loops and makes sensitivity analyses reproducible (Taylor, 2008). Relative to earlier volatility comparison papers that prioritized point accuracy or variance forecasts alone (Hansen & Lunde, 2005), the

American Journal of Scholarly Research and Innovation
Volume 04, Issue 01 (2025)
Page No: 287-319
eISSN: 3067-5146
Doi: 10.63125/z8qq6h36
present synthesis underscores that production-grade success rests as much on coverage stability and exceedance independence as on mean-squared error. In volatile, internationally connected markets, that emphasis is consistent with supervisory expectations and with internal model-risk standards that privilege coherent, auditable risk outputs over marginal accuracy gains.

Our MLOps analysis complements the modeling picture by showing that adoption of registries, CI/CD, and canary/shadow rollouts is meaningful but incomplete. This observation dovetails with software-engineering case studies arguing that ML systems require adapted processes for data versioning, lineage, and continuous testing (Amershi et al., 2019). The documented benefit smaller backtest-to-live loss gaps and lower rollback rates when these practices are present matches industry guidance on data validation and pipeline hygiene (Polyzotis et al., 2017)and with large-scale efforts to automate data quality checks as first-class tests (Schelter et al., 2018). Standardized documentation artifacts such as model cards and datasheets also appear in the higher-reliability subset of studies, aligning with proposals to institutionalize usage boundaries, assumptions, and evaluation slices (Gebru et al., 2021; Mitchell et al., 2019). The comparison with earlier engineering literature suggests a practical hierarchy: feature stores and registries reduce configuration drift; CI/CD enforces repeatable builds and evaluations; canary/shadow plus monitoring catches distribution shifts and tail calibration drift before full exposure. Our contribution is to tie those practices quantitatively to forecasting and risk outcomes, showing that operational discipline is not a compliance luxury but a performance multiplier especially when promotion decisions hinge on small but trustworthy gains.

Figure 12: Hybrid Model for Stock Forecasting and Risk in Volatile Markets



American Journal of Scholarly Research and Innovation
Volume 04, Issue 01 (2025)
Page No: 287-319
eISSN: 3067-5146
Doi: 10.63125/z8qq6h36

CONCLUSION

This research concludes that hybrid machine learning systems engineered and *deployed* with production discipline provide reliably better stock price forecasting and risk prediction under volatile, internationally integrated market conditions, transforming modest accuracy gains into materially improved uncertainty control and tail behavior. Synthesizing 120 PRISMA-screened studies, we found steady single-digit reductions in point error (median RMSE −7.8%, MAE −7.1%) and directional gains of roughly +3.6 percentage points, paired with distributional improvements (pinball −6.1%, CRPS −5.4%) and tighter Value-at-Risk coverage at both 1% and 5% levels (roughly halving nominal deviations). These results persist under leakage-safe, rolling evaluation and remain calibrationally robust when volatility rises, even as mean-squared gains compress, which is exactly the trade-off demanded by production risk controls. The evidence isolates design choices that travel well from lab to line: a decomposable statistical baseline to anchor low-frequency structure; a tabular expert to capture cross-feature interactions; and a sequence/attention expert to model temporal context combined through *dynamic, regime-aware weighting* that adds ≈+2.9% extra RMSE reduction over static weights overall and ≈+3.5% in turbulent windows. Measurement choices magnify these effects: realized-volatility features cut pinball by ≈3.1% and tighten 1% VaR by ≈0.4 percentage points; portable microstructure proxies boost short-horizon direction by ≈+2.8 percentage points; and lightweight range-based estimators deliver ≈−1.9% RMSE at negligible compute cost. Yet the decisive differentiator is operational discipline. While 70.0% of studies implemented full pipelines and 52.5% instrumented drift or exceedance monitoring, only 47.5% used registries, 34.2% documented CI/CD, 24.2% used canary/shadow rollouts, and 18.3% declared explicit SLOs; where registry+CI/CD+canary co-occurred, the backtest-to-live loss gap shrank to ≈1.8% and rollback rates were markedly lower, demonstrating that reproducibility and gated promotion are *performance multipliers*, not mere governance niceties. The resulting deployment blueprint is clear and actionable: (i) a measurement layer producing realized and range-based volatility plus portable microstructure features via idempotent, time-aligned transformations; (ii) a hybrid modeling layer with statistical, tabular, and sequence experts; (iii) a meta-learner that sets dynamic, regime-aware weights from rolling residuals; (iv) a single probabilistic head emitting quantiles for forecasting and risk dashboards; and (v) a platform layer feature store, model registry, CI/CD gates, canary/shadow exposure, drift and exceedance monitoring, and rollback playbooks tied to SLOs for latency, throughput, freshness, and calibration. Limitations in the literature sparse latency and cost reporting, heterogeneous baselines, occasional gaps in conditional-coverage diagnostics reflect documentation more than feasibility and indicate where standardization would further reduce evaluation-to-production drift. Overall, the synthesis supports a practical claim: hybrids, when implemented with disciplined data contracts and promotion controls, reliably convert incremental accuracy into robust probabilistic calibration and risk adequacy, yielding forecasting and VaR/ES services that are reproducible, explainable, and safe to operate at the speed and stress levels of modern equity markets.

RCOMMENDATIONS

Building on the evidence from this review, we recommend that organizations treat hybrid forecasting-and-risk solutions as engineered, versioned products rather than one-off models, and implement them through a five-layer blueprint that converts modest accuracy gains into dependable calibration under volatility. First, harden the measurement layer: enforce idempotent, time-zone-aligned OHLCV pipelines with corporate-action adjustments, snapshot raw and derived fields, and compute low-cost range estimators (e.g., high–low, open–close composites) alongside realized-volatility measures where intraday data exist; standardize microstructure proxies (order imbalance, implicit spread) as portable features so the same artifact can serve across venues. Second, institutionalize leakage control: require rolling-origin (walk-forward) evaluation; freeze scalers and encoders on training windows; adopt purge/embargo around labels; and forbid random K-fold splits for any promotion decision. Third, assemble the hybrid modeling layer with components that specialize and complement: a decomposable statistical baseline to anchor trend/seasonality, a tabular expert (gradient-boosted trees or equivalent) to capture cross-feature interactions, and a sequence/attention expert for temporal dependencies; train a meta-learner that assigns dynamic, regime-aware weights using out-of-fold residuals, and keep ablation reports (−sequence, −tabular, −volatility filter) as a standing artifact to justify each member's inclusion. Fourth, make probabilistic output first-class: predict quantiles directly for multiple horizons and drive VaR/ES from the same

head; wrap the model with lightweight calibration (e.g., rolling conformal or expectile alignment) to maintain target coverage when distributions shift. Fifth, elevate MLOps from optional to mandatory: register every model, feature graph, and config in a model registry; gate releases with CI/CD that replays the entire walk-forward suite and nesting-aware tests against a strong baseline; use canary or shadow exposure with explicit SLOs e.g., sub-25 ms median inference, freshness within one bar, and VaR deviation thresholds of ±0.5 pp at 1% and ±0.7 pp at 5% and declare rollback playbooks that specify when to widen intervals, freeze weights, or fall back to a volatility-only baseline. Add observability tuned to risk: schema and outlier checks at ingestion, live drift monitors on covariates and residuals, exceedance dashboards for VaR/ES with independence kernels, and alert budgets tied to business loss limits; require post-incident reviews that update tests and playbooks. For explainability and governance, publish model cards and dataset datasheets, expose SHAP or permutation attributions for the tabular expert, and localized explanations or example-based diagnostics for the sequence component; make all artifacts reproducible via pinned environments and checksums. Finally, run operations with cost discipline: default to CPU inference with batch pre-computation for the heaviest horizons, reserve GPUs for retraining bursts, schedule daily/weekly retrains with regime-triggered fast paths, and enforce a "no-promote without backtest parity" rule that compares live residuals to backtest distributions for the first week of traffic. Teams that execute this blueprint measurement rigor, leakage-safe evaluation, adaptive hybridization, calibrated probabilistic outputs, and production-grade MLOps will deploy hybrids that are auditable, resilient, and economically useful when markets are most volatile.

REFERENCES

[1]. Abdullah Al, M., Md Masud, K., Mohammad, M., & Hosne Ara, M. (2024). Behavioral Factors in Loan Default Prediction A Literature Review On Psychological And Socioeconomic Risk Indicators. *American Journal of Advanced Technology and Engineering Solutions*, *4*(01), 43-70. https://doi.org/10.63125/0jwtbn29

[2]. 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

[3]. Acerbi, C., & Tasche, D. (2002). Expected shortfall: A natural coherent alternative to value at risk. *Journal of Banking & Finance*, *26*(7), 1507-1533. https://doi.org/https://doi.org/10.1016/S0378-4266(02)00283-2

[4]. 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

[5]. Amershi, S., Begel, A., Bird, C., DeLine, R., Gall, H., Kamar, E., Nagappan, N., Nushi, B., & Zimmermann, T. (2019). Software engineering for machine learning: A case study. 2019 IEEE/ACM 41st International Conference on Software Engineering: Software Engineering in Practice (ICSE-SEIP),

[6]. Amisano, G., & Giacomini, R. (2007). Comparing density forecasts via weighted likelihood ratio tests. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *69*(4), 627-646. https://doi.org/https://doi.org/10.1111/j.1467-9868.2007.00653.x

[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]. Anika Jahan, M., Md Soyeb, R., & Tahmina Akter, R. (2025). Strategic Use Of Engagement Marketing in Digital Platforms: A Focused Analysis Of Roi And Consumer Psychology. *Journal of Sustainable Development and Policy*, *1*(01), 170-197. https://doi.org/10.63125/hm96p734

[9]. Antweiler, W., & Frank, M. Z. (2004). Is all that talk just noise? The information content of internet stock message boards. *The Journal of Finance*, *59*(3), 1259-1294. https://doi.org/https://doi.org/10.1111/j.1540-6261.2004.00662.x

[10]. 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

[11]. Athey, S., Tibshirani, J., & Wager, S. (2019). Generalized random forests. *The Annals of Statistics*, *47*(2), 1148-1178. https://doi.org/https://doi.org/10.1214/19-AOS1709

[12]. 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

[13]. Bai, S., Kolter, J. Z., & Koltun, V. (2018). An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. *arXiv*. https://doi.org/https://doi.org/10.48550/arXiv.1803.01271

[14]. Ballings, M., Van den Poel, D., Hespeels, N., & Gryp, R. (2015). Evaluating multiple classifiers for stock price direction prediction. *Expert Systems with Applications*, *42*(20), 7046-7056. https://doi.org/https://doi.org/10.1016/j.eswa.2015.05.013

[15]. 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
[16]. Barndorff-Nielsen, O. E., Hansen, P. R., Lunde, A., & Shephard, N. (2008). Designing realized kernels to measure ex post variation of equity prices. *Econometrica*, *76*(6), 1481-1536. https://doi.org/https://doi.org/10.3982/ECTA6493
[17]. Barndorff-Nielsen, O. E., & Shephard, N. (2002). Econometric analysis of realized volatility and its use in estimating stochastic volatility models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *64*(2), 253-280. https://doi.org/https://doi.org/10.1111/1467-9868.00342
[18]. Bates, J. M., & Granger, C. W. J. (1969). The combination of forecasts. *Operational Research Quarterly*, *20*(4), 451-468. https://doi.org/https://doi.org/10.2307/3008764
[19]. Baylor, D., & et al. (2017). TFX: A TensorFlow-Based Production-Scale Machine Learning Platform. Proceedings of KDD,
[20]. 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
[21]. 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
[22]. Berkowitz, J., Christoffersen, P., & Pelletier, D. (2011). Evaluating value-at-risk models with desk-level data. *Management Science*, *57*(12), 2213-2227. https://doi.org/https://doi.org/10.1287/mnsc.1110.1376
[23]. 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
[24]. Breiman, L. (1996). Bagging predictors. *Machine Learning*, *24*(2), 123-140. https://doi.org/https://doi.org/10.1023/A:1018054314350
[25]. Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5-32. https://doi.org/https://doi.org/10.1023/A:1010933404324
[26]. Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. Proceedings of KDD,
[27]. Christoffersen, P. F. (1998). Evaluating interval forecasts. *International Economic Review*, *39*(4), 841-862. https://doi.org/https://doi.org/10.2307/2527341
[28]. Clark, T. E., & McCracken, M. W. (2001). Tests of equal forecast accuracy and encompassing for nested models. *Journal of Econometrics*, *105*(1), 85-110. https://doi.org/https://doi.org/10.1016/S0304-4076(01)00058-0
[29]. Clark, T. E., & West, K. D. (2007). Approximately unbiased tests for equal predictive accuracy in nested models. *Journal of Business & Economic Statistics*, *25*(4), 389-403. https://doi.org/https://doi.org/10.1198/073500107000000330
[30]. Clemen, R. T. (1989). Combining forecasts: A review and annotated bibliography. *International Journal of Forecasting*, *5*(4), 559-583. https://doi.org/https://doi.org/10.1016/0169-2070(89)90012-5
[31]. Corradi, V., & Swanson, N. R. (2006). Predictive density evaluation. *Journal of Econometrics*, *135*(1–2), 125-154. https://doi.org/https://doi.org/10.1016/j.jeconom.2005.07.001
[32]. 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
[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]. Diebold, F. X., & Yilmaz, K. (2012). Better to give than to receive: Predictive directional measurement of volatility spillovers. *International Journal of Forecasting*, *28*(1), 57-66. https://doi.org/https://doi.org/10.1016/j.ijforecast.2011.02.006
[35]. Engle, R. (2002). Dynamic conditional correlation: A simple class of multivariate GARCH models. *Journal of Business & Economic Statistics*, *20*(3), 339-350. https://doi.org/https://doi.org/10.1198/073500102288618487
[36]. Engle, R. F. (1982). Autoregressive conditional heteroscedasticity with estimates of the variance of U.K. inflation. *Econometrica*, *50*(4), 987-1007. https://doi.org/https://doi.org/10.2307/1912773
[37]. 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
[38]. Fama, E. F. (1970). Efficient capital markets: A review of theory and empirical work. *The Journal of Finance*, *25*(2), 383-417. https://doi.org/https://doi.org/10.1111/j.1540-6261.1970.tb00518.x
[39]. 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
[40]. Fissler, T., & Ziegel, J. F. (2016). Higher order elicitability and Osband's principle. *The Annals of Statistics*, *44*(4), 1680-1707. https://doi.org/https://doi.org/10.1214/16-AOS1439

[41]. Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, *29*(5), 1189-1232. https://doi.org/https://doi.org/10.1214/aos/1013203451
[42]. Friedman, J. H. (2002). Stochastic gradient boosting. *Computational Statistics & Data Analysis*, *38*(4), 367-378. https://doi.org/https://doi.org/10.1016/S0167-9473(01)00065-2
[43]. Gama, J., Žliobaitė, I., Bifet, A., Pechenizkiy, M., & Bouchachia, A. (2014). A survey on concept drift adaptation. *ACM Computing Surveys*, *46*(4), 44. https://doi.org/https://doi.org/10.1145/2523813
[44]. Garman, M. B., & Klass, M. J. (1980). On the estimation of security price volatilities from historical data. *Journal of Business*, *53*(1), 67-78. https://doi.org/https://doi.org/10.1086/296093
[45]. Gebru, T., Morgenstern, J., Vecchione, B., Wortman Vaughan, J., Wallach, H., Daumé III, H., & Crawford, K. (2021). Datasheets for datasets. *Communications of the ACM*, *64*(12), 86-92. https://doi.org/https://doi.org/10.1145/3458723
[46]. Geweke, J., & Amisano, G. (2011). Optimal prediction pools. *Journal of Econometrics*, *164*(1), 130-141. https://doi.org/https://doi.org/10.1016/j.jeconom.2011.02.016
[47]. Giacomini, R., & Rossi, B. (2010). Forecast comparisons in unstable environments. *Journal of Econometrics*, *158*(1), 58-73. https://doi.org/https://doi.org/10.1016/j.jeconom.2010.03.022
[48]. 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
[49]. Giot, P., & Laurent, S. (2004). Modelling daily value-at-risk using realized volatility and ARCH type models. *Journal of Empirical Finance*, *11*(3), 379-398. https://doi.org/https://doi.org/10.1016/j.jempfin.2003.03.006
[50]. Glosten, L. R., Jagannathan, R., & Runkle, D. E. (1993). On the relation between the expected value and the volatility of the nominal excess return on stocks. *The Journal of Finance*, *48*(5), 1779-1801. https://doi.org/https://doi.org/10.1111/j.1540-6261.1993.tb05128.x
[51]. 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
[52]. 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
[53]. Goyal, A., & Welch, I. (2008). A comprehensive look at the empirical performance of equity premium prediction. *The Review of Financial Studies*, *21*(4), 1455-1508. https://doi.org/https://doi.org/10.1093/rfs/hhm034
[54]. Gu, S., Kelly, B., & Xiu, D. (2020). Empirical asset pricing via machine learning. *The Review of Financial Studies*, *33*(5), 2223-2273. https://doi.org/https://doi.org/10.1093/rfs/hhaa009
[55]. 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
[56]. Hansen, P. R. (2005). A test for superior predictive ability. *Journal of Business & Economic Statistics*, *23*(4), 365-380. https://doi.org/https://doi.org/10.1198/073500104000000751
[57]. 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
[58]. 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
[59]. Hansen, P. R., Lunde, A., & Nason, J. M. (2011). The model confidence set. *Econometrica*, *79*(2), 453-497. https://doi.org/https://doi.org/10.3982/ECTA5771
[60]. Hasbrouck, J. (1991). Measuring the information content of stock trades. *The Journal of Finance*, *46*(1), 179-207. https://doi.org/https://doi.org/10.1111/j.1540-6261.1991.tb04636.x
[61]. Hendershott, T., Jones, C. M., & Menkveld, A. J. (2011). Does algorithmic trading improve liquidity? *The Journal of Finance*, *66*(1), 1-33. https://doi.org/https://doi.org/10.1111/j.1540-6261.2010.01624.x
[62]. 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
[63]. 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
[64]. 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
[65]. Hyndman, R. J., & Khandakar, Y. (2008). Automatic time series forecasting: The forecast package for R. *Journal of Statistical Software*, *27*(3), 1-22. https://doi.org/https://doi.org/10.18637/jss.v027.i03
[66]. 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

[67]. 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

[68]. 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

[69]. 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

[70]. Jordan, M. I., & Jacobs, R. A. (1994). Hierarchical mixtures of experts and the EM algorithm. *Neural Computation*, 6(2), 181-214. https://doi.org/https://doi.org/10.1162/neco.1994.6.2.181

[71]. Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., & Liu, T.-Y. (2017). LightGBM: A highly efficient gradient boosting decision tree. Advances in Neural Information Processing Systems,

[72]. Khan, A. S., Akter, M., Enni, M. A., & Khan, S. F. (2025). An in silico approach for the identification of detrimental missense SNPs and their potential impacts on human CRY2 protein. *Journal of Bangladesh Academy of Sciences*, *49*(1), 57-72. https://doi.org/10.3329/jbas.v49i1.71914

[73]. 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

[74]. Koenker, R., & Bassett, G., Jr. (1978). Regression quantiles. *Econometrica*, *46*(1), 33-50. https://doi.org/https://doi.org/10.2307/1913643

[75]. 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.2016.10.089

[76]. Kuester, K., Mittnik, S., & Paolella, M. S. (2006). Value-at-risk prediction: A comparison of alternative strategies. *Journal of Financial Econometrics*, *4*(1), 53-89. https://doi.org/https://doi.org/10.1093/jjfinec/nbj002

[77]. 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

[78]. Lim, B., Arik, S. O., Loeff, N., & Pfister, T. (2021). Temporal Fusion Transformers for interpretable multi-horizon time series forecasting. *International Journal of Forecasting*, *37*(4), 1748-1764. https://doi.org/https://doi.org/10.1016/j.ijforecast.2021.03.012

[79]. Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems,

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

[81]. Makridakis, S., Spiliotis, E., & Assimakopoulos, V. (2020). The M4 competition: 100,000 time series and 61 forecasting methods. *International Journal of Forecasting*, *36*(1), 54-74. https://doi.org/https://doi.org/10.1016/j.ijforecast.2019.04.014

[82]. 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

[83]. Mansura Akter, E. (2025). Bioinformatics-Driven Approaches in Public Health Genomics: A Review Of Computational SNP And Mutation Analysis. *International Journal of Scientific Interdisciplinary Research*, *6*(1), 88-118. https://doi.org/10.63125/e6pxkn12

[84]. 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

[85]. 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

[86]. 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

[87]. Md Arafat, S., Md Imran, K., Hasib, A., Md Jobayer Ibne, S., & Md Sanjid, K. (2025). Investigating Key Attributes for Circular Economy Implementation In Manufacturing Supply Chains: Impacts On The Triple Bottom Line. *Review of Applied Science and Technology*, *4*(02), 145-175. https://doi.org/10.63125/fnsy0e41

[88]. 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

[89]. 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

[90]. Md Atiqur Rahman, K., Md Abdur, R., Niger, S., & Mst Shamima, A. (2025). Development Of a Fog Computing-Based Real-Time Flood Prediction And Early Warning System Using Machine Learning And Remote Sensing Data. *Journal of Sustainable Development and Policy*, *1*(01), 144-169. https://doi.org/10.63125/6y0qwr92

[91]. 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

[92]. 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

[93]. 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

[94]. Md Jakaria, T., Md, A., Zayadul, H., & Emdadul, H. (2025). Advances In High-Efficiency Solar Photovoltaic Materials: A Comprehensive Review of Perovskite And Tandem Cell Technologies. *American Journal of Advanced Technology and Engineering Solutions*, *1*(01), 201-225. https://doi.org/10.63125/5amnvb37

[95]. 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

[96]. Md Masud, K., Mohammad, M., & Hosne Ara, M. (2023). Credit decision automation in commercial banks: a review of AI and predictive analytics in loan assessment. *American Journal of Interdisciplinary Studies*, *4*(04), 01-26. https://doi.org/10.63125/1hh4q770

[97]. 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

[98]. Md Masud, K., Sazzad, I., Mohammad, M., & Noor Alam, S. (2025). Digitization In Retail Banking: A Review of Customer Engagement And Financial Product Adoption In South Asia. *ASRC Procedia: Global Perspectives in Science and Scholarship*, *1*(01), 42-46. https://doi.org/10.63125/cv50rf30

[99]. Md, N., Golam Qibria, L., Abdur Razzak, C., & Khan, M. A. M. (2025). Predictive Maintenance In Power Transformers: A Systematic Review Of AI And IOT Applications. *ASRC Procedia: Global Perspectives in Science and Scholarship*, *1*(01), 34-47. https://doi.org/10.63125/r72yd809

[100]. Md Nazrul Islam, K., & Debashish, G. (2025). Cybercrime and contractual liability: a systematic review of legal precedents and risk mitigation frameworks. *Journal of Sustainable Development and Policy*, *1*(01), 01-24. https://doi.org/10.63125/x3cd4413

[101]. Md Nazrul Islam, K., & Ishtiaque, A. (2025). A systematic review of judicial reforms and legal access strategies in the age of cybercrime and digital evidence. *International Journal of Scientific Interdisciplinary Research*, *5*(2), 01-29. https://doi.org/10.63125/96ex9767

[102]. 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

[103]. 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

[104]. 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

[105]. Md Takbir Hossen, S., Abdullah Al, M., Siful, I., & Md Mostafizur, R. (2025). Transformative applications of ai in emerging technology sectors: a comprehensive meta-analytical review of use cases in healthcare, retail, and cybersecurity. *ASRC Procedia: Global Perspectives in Science and Scholarship*, *1*(01), 121-141. https://doi.org/10.63125/45zpb481

[106]. 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

[107]. 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

[108]. 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

[109]. Md Tawfiqul, I. (2025). Adversarial Defence Mechanisms In Neural Networks For ICS Fault Tolerance: A Comparative Analysis. *ASRC Procedia: Global Perspectives in Science and Scholarship*, *1*(01), 404-431. https://doi.org/10.63125/xrp7be57

[110]. 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

[111]. 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

[112]. 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

[113]. Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I. D., & Gebru, T. (2019). Model cards for model reporting. Proceedings of the Conference on Fairness, Accountability, and Transparency,

[114]. Montero-Manso, P., Athanasopoulos, G., Hyndman, R. J., & Talagala, T. S. (2020). FFORMA: Feature-based forecast model averaging. *International Journal of Forecasting*, *36*(1), 86-92. https://doi.org/https://doi.org/10.1016/j.ijforecast.2019.02.011

[115]. 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

[116]. 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

[117]. Mullainathan, S., & Spiess, J. (2017). Machine learning: An applied econometric approach. *Journal of Economic Perspectives*, *31*(2), 87-106. https://doi.org/https://doi.org/10.1257/jep.31.2.87

[118]. Nelson, D. B. (1991). Conditional heteroskedasticity in asset returns: A new approach. *Journal of Econometrics*, *45*(1-2), 347-364. https://doi.org/https://doi.org/10.1016/0304-4076(91)90201-8

[119]. Parkinson, M. (1980). The extreme value method for estimating the variance of the stock price. *Journal of Business*, *53*(1), 61-65. https://doi.org/https://doi.org/10.1086/296071

[120]. Patel, J., Shah, S., Thakkar, P., & Kotecha, K. (2015). Predicting stock market index movement using machine learning techniques. *Expert Systems with Applications*, *42*(20), 2162-2172. https://doi.org/https://doi.org/10.1016/j.eswa.2015.06.016

[121]. Patton, A. J. (2011). Volatility forecast comparison using imperfect volatility proxies. *Journal of Econometrics*, *160*(1), 246-256. https://doi.org/https://doi.org/10.1016/j.jeconom.2010.03.034

[122]. Polyzotis, N., Roy, S., Whang, S. E., & Zinkevich, M. (2017). Data management challenges in production machine learning. Proceedings of the 2017 ACM International Conference on Management of Data (SIGMOD '17),

[123]. Preis, T., Moat, H. S., & Stanley, H. E. (2013). Quantifying trading behavior in financial markets using Google Trends. *Scientific Reports*, *3*, 1684. https://doi.org/https://doi.org/10.1038/srep01684

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

[125]. Raftery, A. E., Kárný, M., & Ettler, P. (2005). Online prediction under model uncertainty via dynamic model averaging. *Technometrics*, *49*(2), 138-154. https://doi.org/https://doi.org/10.1198/004017005000000014

[126]. Rangapuram, S. S., Seeger, M. W., Gasthaus, J., Stella, L., Wang, Y., & Januschowski, T. (2018). Deep state space models for time series forecasting. Advances in Neural Information Processing Systems (NeurIPS 2018),

[127]. Rapach, D. E., Strauss, J. K., & Zhou, G. (2010). Out-of-sample equity premium prediction: Combination forecasts and links to the real economy. *Journal of Financial Economics*, *96*(2), 217-231. https://doi.org/https://doi.org/10.1016/j.jfineco.2010.02.013

[128]. 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
[129]. 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
[130]. Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?": Explaining the predictions of any classifier. Proceedings of KDD,
[131]. Roll, R. (1984). A simple implicit measure of the effective bid–ask spread in an efficient market. *The Journal of Finance*, *39*(4), 1127-1139. https://doi.org/https://doi.org/10.1111/j.1540-6261.1984.tb03897.x
[132]. Romano, J. P., & Wolf, M. (2005). Stepwise multiple testing as formalized data snooping. *Econometrica*, *73*(4), 1237-1282. https://doi.org/https://doi.org/10.1111/j.1468-0262.2005.00615.x
[133]. Salinas, D., Flunkert, V., Gasthaus, J., & Januschowski, T. (2020). DeepAR: Probabilistic forecasting with autoregressive recurrent networks. *International Journal of Forecasting*, *36*(3), 1181-1191. https://doi.org/https://doi.org/10.1016/j.ijforecast.2019.07.001
[134]. 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
[135]. 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
[136]. Sazzad, I. (2025a). Public Finance and Policy Effectiveness A Review Of Participatory Budgeting In Local Governance Systems. *Journal of Sustainable Development and Policy*, *1*(01), 115-143. https://doi.org/10.63125/p3p09p46
[137]. Sazzad, I. (2025b). A Systematic Review of Public Budgeting Strategies In Developing Economies: Tools For Transparent Fiscal Governance. *American Journal of Advanced Technology and Engineering Solutions*, *1*(01), 602-635. https://doi.org/10.63125/wm547117
[138]. 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
[139]. Schwert, G. W. (1989). Why does stock market volatility change over time? *The Journal of Finance*, *44*(5), 1115-1153. https://doi.org/https://doi.org/10.1111/j.1540-6261.1989.tb00519.x
[140]. 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
[141]. 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
[142]. Sohel, R., & Md, A. (2022). A Comprehensive Systematic Literature Review on Perovskite Solar Cells: Advancements, Efficiency Optimization, And Commercialization Potential For Next-Generation Photovoltaics. *American Journal of Scholarly Research and Innovation*, *1*(01), 137-185. https://doi.org/10.63125/843z2648
[143]. Stock, J. H., & Watson, M. W. (2002). Macroeconomic forecasting using diffusion indexes. *Econometrica*, *70*(2), 613-646. https://doi.org/https://doi.org/10.1111/1468-0262.00273
[144]. Subrato, S. (2025). Role of management information systems in environmental risk assessment: a systematic review of geographic and ecological applications. *American Journal of Interdisciplinary Studies*, *6*(1), 95–126. https://doi.org/10.63125/k27tnn83
[145]. Subrato, S., & Faria, J. (2025). AI-driven MIS applications in environmental risk monitoring: a systematic review of predictive geographic information systems. *ASRC Procedia: Global Perspectives in Science and Scholarship*, *1*(01), 81-97. https://doi.org/10.63125/pnx77873
[146]. 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
[147]. Tahmina Akter, R. (2025). AI-driven marketing analytics for retail strategy: a systematic review of data-backed campaign optimization. *International Journal of Scientific Interdisciplinary Research*, *6*(1), 28-59. https://doi.org/10.63125/0k4k5585
[148]. Tahmina Akter, R., & Abdur Razzak, C. (2022). The Role Of Artificial Intelligence In Vendor Performance Evaluation Within Digital Retail Supply Chains: A Review Of Strategic Decision-Making Models. *American Journal of Scholarly Research and Innovation*, *1*(01), 220-248. https://doi.org/10.63125/96jj3j86

[149]. 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

[150]. Tahmina Akter, R., Md Arifur, R., & Anika Jahan, M. (2024). Customer relationship management and data-driven decision-making in modern enterprises: a systematic literature review. *American Journal of Advanced Technology and Engineering Solutions*, *4*(04), 57-82. https://doi.org/10.63125/jetvam38

[151]. 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/nbn004

[152]. 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

[153]. Tetlock, P. C. (2007). Giving content to investor sentiment: The role of media in the stock market. *The Journal of Finance*, *62*(3), 1139-1168. https://doi.org/https://doi.org/10.1111/j.1540-6261.2007.01219.x

[154]. Timmermann, A. (2006). Forecast combinations. *Journal of Econometrics*, *135*(1-2), 197-228. https://doi.org/https://doi.org/10.1016/j.jeconom.2005.07.005

[155]. 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

[156]. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *arXiv*. https://doi.org/https://doi.org/10.48550/arXiv.1706.03762

[157]. Wen, R., Torkkola, K., Narayanaswamy, B., & Madeka, D. (2017). A multi-horizon quantile recurrent forecaster. *arXiv*. https://doi.org/https://doi.org/10.48550/arXiv.1711.11053

[158]. West, K. D. (1996). Asymptotic inference about predictive ability. *Econometrica*, *64*(5), 1067-1084. https://doi.org/https://doi.org/10.2307/2171956

[159]. White, H. (2000). A reality check for data snooping. *Econometrica*, *68*(5), 1097-1126. https://doi.org/https://doi.org/10.1111/1468-0262.00078

[160]. 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

[161]. Wu, H., Xu, J., Wang, J., & Long, M. (2021). Autoformer: Decomposition transformers with auto-correlation for long-term series forecasting. *arXiv*. https://doi.org/https://doi.org/10.48550/arXiv.2106.13008

[162]. Yang, D., & Zhang, Q. (2000). Drift-independent volatility estimation based on high, low, open, and close prices. *Journal of Business*, *73*(3), 477-492. https://doi.org/https://doi.org/10.1086/209650

[163]. Zaharia, M., Chen, A., Davidson, A., & et al. (2018). Accelerating the machine learning lifecycle with MLflow. Proceedings of the 2018 International Conference on Management of Data (SIGMOD),

[164]. 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

[165]. 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

[166]. Zhou, H., Zhang, S., Peng, J., Zhang, S., Li, J., Xiong, H., & Zhang, W. (2021). Informer: Beyond efficient Transformer for long sequence time-series forecasting. Proceedings of the AAAI Conference on Artificial Intelligence,