
Research article

Uncertainty calibrated volatility forecasting through regime detection and adaptive ensemble modeling

Khalid Jeaab^{1,*}, Smaaine Ouaharahe² and Youness Saoudi³

¹ Sultan Moulay Slimane University, Economics and Management Laboratory, Khouribga, Morocco

² Ibn Tofail University, Organization Economics and Management Laboratory, Kenitra, Morocco

³ Ibn Tofail University, Advanced Systems Engineering Laboratory, Kenitra, Morocco

* **Correspondence:** Email: jeaab.khalid.feg@usms.ac.ma.

Abstract: Accurate volatility forecasting, which is critical for financial risk management, has been challenged by classic econometric models and recent machine learning algorithms that lack reliable uncertainty quantification. The regime-adaptive deep ensemble with uncertainty quantification solution combines bidirectional long short-term memory (Bi-LSTM) for regime identification and bootstrap uncertainty quantification with models such as the heterogeneous autoregressive model for realized volatility (HAR-RV), random forest, and light gradient-boosting machine (LightGBM). It also includes triple-adaptive ensemble weighting. A 10-year investigation of S&P 500 data showed a 55.6% improvement in RMSE compared to the heterogeneous autoregressive model for realized volatility (HAR-RV) baseline and an R^2 increase from 0.602 to 0.922. The approach demonstrated confirmed 95% prediction intervals, outstanding adaptability, and performance across several regimes. Cross-asset validation across five asset classes (S&P 500, NASDAQ-100, Russell 2000, EUR/USD, and Gold) confirms consistent root mean square error (RMSE) improvements of 47–55% over HAR-RV and R^2 above 0.89 across all assets, demonstrating robust generalizability. Economic evaluations show a 94.6% reduction in portfolio volatility and an enhanced Sharpe ratio. This framework outperforms previous methods by providing high accuracy, well-calibrated uncertainty estimates, and a thorough economic analysis.

Keywords: volatility forecasting; ensemble learning; uncertainty quantification; bootstrap methods; regime detection; financial econometrics

JEL Codes: C53; C58; C45; C22; G17; G11

1. Introduction

Financial market volatility, the degree of price change over time, is a critical factor in modern finance, influencing risk management, portfolio optimization, option pricing, and regulatory capital calculations. Unlike asset returns, which are mainly unpredictable, volatility demonstrates significant persistence and forecastability (Andersen et al., 2003; Bollerslev, 1986), making precise volatility forecasting possible and economically advantageous.

The volatility revolution established high-frequency data as the empirical baseline, allowing for model-free volatility estimates (Andersen et al., 2001; Barndorff-Nielsen and Shephard, 2002). The heterogeneous autoregressive model of realized volatility (HAR-RV) (Corsi, 2009) became the leading forecasting model due to its simplicity, interpretability, and strong empirical performance. However, the heterogeneous autoregressive model for realized volatility (HAR-RV)'s linear form restricts its capacity to capture complex nonlinearities and regime-dependent dynamics common in financial markets.

Recent machine learning developments provide potential options. Tree-based approaches, neural networks, and gradient boosting outperform classical models in terms of prediction accuracy, with root mean square error (RMSE) reductions ranging from 20–30% (Christensen et al., 2023). However, these improvements come at a hefty cost: (1) Lack of uncertainty quantification—most research gives only point forecasts without prediction intervals. (2) Poor calibration—intervals often show consistent under- or over-coverage. (3): Static models cannot adjust to changing market conditions. (4): Traditional evaluations prioritize point accuracy metrics (root mean square error (RMSE), mean absolute error (MAE)) above calibration quality, cross-regime stability, and economic value.

Furthermore, uncertainty quantification is not only a theoretical nicety but also a real requirement. Risk management applications require reliable prediction intervals for Value-at-Risk (VaR) assessments. Regulatory frameworks (Basel III) need statistically reliable risk models. Portfolio optimization with parameter uncertainty necessitates well-calibrated second moments. However, few volatility forecasting studies provide rigorous uncertainty quantification and confirm calibration quality using statistical tests (Kupiec, 1995; Christoffersen, 1998).

1.1. Proposed solution and contributions

We present RADE-UQ (regime-adaptive deep ensemble with uncertainty quantification), a comprehensive architecture that addresses crucial gaps through five interconnected components: (1) The regime detection network, a bidirectional long short-term memory (Bi-LSTM) design with hierarchical attention, detects latent volatility regimes without making parametric assumptions. (2) Bootstrap uncertainty quantification - Each basic model (HAR-RV, random forest, light gradient boosting machine (LightGBM)) is trained with 50 iterations for a principled uncertainty estimate (3). The heterogeneous ensemble combines econometric and machine learning models for complementing strengths. (4) Triple-adaptive weighting. Ensemble weights adapt through regime-conditional, uncertainty-based, performance-adaptive processes. (5) Comprehensive evaluation includes more than 15 innovative criteria evaluating calibration quality, uncertainty decomposition, regime-specific performance, and economic value.

RADE-UQ uses 10 years of S&P 500 data (2767 observations, 2014–2024) and six macroeconomic variables to make four key contributions:

C1: State-of-the-art predictive performance. RADE-UQ obtains an RMSE of 0.0138, demonstrating

55.6% improvement over the HAR-RV baseline (RMSE: 0.0310). The coefficient of determination (R^2) increases from 0.602 to 0.922. This outperforms current machine learning benchmarks (Christensen et al., 2023) while being computationally efficient.

C2: Rigorous uncertainty quantification with validation. RADE-UQ not only delivers point forecasts but also statistically valid prediction intervals. Kupiec's unconditional coverage test validates the post-calibration empirical coverage at 94.97% (target: 95%) ($p = 0.923$). An interval width quality of 0.962 indicates remarkable adaptive capacity—intervals extend considerably under high volatility and compress during calm times. This is the first volatility forecasting research to successfully achieve high point accuracy, well-calibrated intervals, statistical validation, and adaptive interval widths.

C3: Comprehensive evaluation framework. We offer an 18-metric framework that includes six metrics for calibration quality, three for uncertainty decomposition, five for regime-specific performance, and four for economic value. Key advances include interval width quality (measuring adaptive capability), regime stability index (cross-regime consistency: 0.856), and uncertainty composition ratio (epistemic vs. aleatoric decomposition: 36.2%/63.8%).

C4: Practical economic value. An economic value study confirms that forecast improvements lead to significant risk management gains, with portfolio variance reduced by 94.6%, a VaR hit rate of 1.32% (target: 1%), and a Sharpe ratio improvement of +0.015.

The remainder of this article is as follows. Section 2 discusses relevant literature. Section 3 describes the RADE-UQ framework design. Section 4 discusses data and experimental techniques. Section 5 provides thorough empirical results. Section 6 analyzes the findings and considers their limits. Section 7 summarizes practical consequences and future research directions.

2. Related work

2.1. Volatility forecasting: evolution and machine learning

Volatility modeling traces back to the foundational autoregressive conditional heteroskedasticity (ARCH) framework (Bollerslev, 1986), which identified volatility clustering. Generalized autoregressive conditional heteroskedasticity (GARCH) extensions ruled for decades, providing both parsimony and theoretical elegance. However, they (Hansen and Lunde, 2005) proved that basic historical averages typically beat complicated GARCH specifications out-of-sample, which calls model-based techniques into doubt.

The realized volatility revolution shifted the field by utilizing high-frequency data for model-free volatility measurement, allowing for more powerful forecast evaluation. The heterogeneous autoregressive (HAR) model, proposed by (Corsi, 2009), incorporates daily, weekly, and monthly realized volatility lags to capture The empirical success of HAR-RV is attributed to its ability to approximate long-memory operations using a simple AR structure.

Machine learning challenges HAR-RV domination. Christensen (2023) conducted a comprehensive evaluation of tree-based approaches (random forest, gradient boosting) versus classical models, finding a 25–30% RMSE decrease. Neural network techniques have mixed results; recurrent structures enhance accuracy somewhat but at significant computational cost.

However, as Christensen (2023) noted, small accuracy increases from complicated models may not translate into economic advantages when transaction costs and estimation uncertainty are included.

This finding inspired our ensemble strategy, which combines interpretable benchmarks (HAR-RV) with flexible learners (random forest, LightGBM) to balance accuracy, efficiency, and resilience. Critically, most research offers only point projections without uncertainty estimates. Our main methodological contribution.

2.2. Ensemble learning and uncertainty quantification

The concept of ensemble learning, which combines many models to improve performance, originated with Breiman's (1996) bootstrap aggregation (bagging). The major takeaway was that averaging predictions from models trained on bootstrap samples decreases variance while retaining bias. Random Forests (Breiman, 2001) improve bagging with decorrelated trees, resulting in impressive empirical results.

The financial forecasting literature is increasingly using ensembles; however, most applications use homogeneous ensembles (e.g., several neural networks) rather than diverse combinations. According to Zhou (2012), heterogeneous ensembles, which combine model families, enhance diversity and robustness. To enhance complementarity, our approach incorporates linear (HAR-RV), tree-based (random forest), and gradient boosting (LightGBM) models.

Despite its practical relevance, uncertainty quantification is underexplored in financial forecasting. Gneiting and Raftery (2007) established theoretical foundations through suitable scoring rules, validating the evaluation of both point predictions (RMSE, MAE) and probabilistic forecasts. Practical uncertainty quantification often uses parametric or bootstrap methodologies. Bootstrap approaches, which use resampling to build empirical distributions, are versatile yet computationally costly (Breiman, 1996).

Unfortunately, most research provides uncertainty estimates without validation. Kupiec (1995) created formal tests for interval forecast accuracy by comparing empirical coverage rates to nominal levels. Christoffersen (1998) expanded this paradigm to assess unconditional coverage and independence. We use these experiments to systematically evaluate our uncertainty estimates, which is a feature missing from most machine learning (ML) forecasting research.

2.3. Regime switching and economic value

Financial time series display regime-dependent dynamics, with structural characteristics changing throughout market situations. Hamilton (1989) proposed Markov-switching models that allow parameters to change among latent regimes. Although theoretically interesting, Markov-switching models have practical limitations, including strong parametric assumptions, computational complexity, and difficulties integrating high-dimensional feature spaces.

Machine learning provides alternate regime identification methodologies. Neural networks with attention mechanisms (Bahdanau et al., 2015; Vaswani et al., 2017) learn to focus on key temporal patterns without explicit regime modeling. Recurrent architectures (Hochreiter and Schmidhuber, 1997; Cho et al., 2014) capture long-range relationships typical of persistent regime structures. Our regime detection network includes: Bi-LSTM for temporal modeling, using hierarchical attention for regime-dependent feature relevance.

Statistical correctness does not always translate into economic benefit. Fleming (2001) pioneered the economic value concept by evaluating volatility forecasts through utility benefits for mean-variance investors, demonstrating that even minor RMSE increases create significant utility gains when forecasts influence portfolio allocation. Fleming (2003) extended this technique to realized volatility projections,

revealing economically substantial utility improvements.

The literature focuses on the alignment of assessment measures with end-user objectives. For risk management applications, well-calibrated prediction intervals may outperform modest point forecast gains. VaR breaches result in regulatory capital fines that considerably surpass usual prediction mistake costs. Our complete evaluation approach considers both statistical and economic features, including portfolio variance reduction, VaR calibration, Sharpe ratio improvements, and realized utility benefits.

2.4. Positioning and advance

RADE-UQ is a one-of-a-kind combination of heterogeneous ensemble learning, bootstrap uncertainty quantification, adaptive regime weighting, and rigorous statistical validation that fills critical gaps in both the ML forecasting literature (lack of UQ and validation) and financial econometrics practice. Our paradigm goes beyond single-model methods by delivering full uncertainty estimates while retaining interpretability via econometric model inclusion. Unlike other ensemble approaches, we incorporate regime detection for dynamic weight adaptation and confirm calibration using formal statistical tests. This multimodal contribution sets new benchmarks for production-ready volatility forecasting systems by addressing accuracy, uncertainty, flexibility, and economic value.

3. Methodology

RADE-UQ consists of five components: (1) data preprocessing and feature engineering, (2) a regime detection network to identify volatility states, (3) bootstrap-based uncertainty quantification for heterogeneous models, (4) triple-adaptive ensemble weighting, and (5) a comprehensive evaluation framework. This section explains the fundamental architecture; further technical specifications are available in Appendix A.

3.1. Problem formulation

Let $\{r_t\}_{t=1}^T$ denote a univariate return series. The realized volatility at time t using M intraday returns is

$$RV_t = \sum_{i=1}^M r_{t,i}^2 \quad (1)$$

Remark on proxy choice. Equation (1) defines the theoretical realized variance using M intraday returns $r_{t,i}$. In our empirical implementation (Section 4.1), we approximate this quantity using the five-day rolling standard deviation of daily log-returns, annualized by $\sqrt{252}$. This daily-frequency proxy is consistent with the HAR-RV benchmark (Corsi, 2009) and the machine learning literature (Christensen et al., 2023), and is robust to proxy noise under the quasi-likelihood (QLIKE) evaluation criterion (Patton, 2011).

We forecast one-day-ahead realized volatility: $\widehat{RV}_{t+1} = f(\mathcal{I}_t)$, where $\mathcal{I}_t$ is the information set at time t . Beyond point forecasts, we seek prediction intervals $[\widehat{RV}_{t+1}^{\text{lower}}, \widehat{RV}_{t+1}^{\text{upper}}]$ satisfying

$$\mathbb{P}\left(RV_{t+1} \in [\widehat{RV}_{t+1}^{\text{lower}}, \widehat{RV}_{t+1}^{\text{upper}}]\right) = 1 - \alpha \quad (2)$$

for confidence level α (typically 0.05 for 95% intervals).

3.2. RADE-UQ architecture overview

Figure 1 provides a complete overview of the RADE-UQ framework architecture, illustrating the five integrated components: feature engineering (42 features across HAR, market, regime, and macroeconomic categories), bootstrap ensemble (50 iterations per model across HAR-RV, random forest, and LightGBM), post-hoc calibration (achieving 95% coverage via multiplier $c^* = 4.4$), and output generation (point forecasts with prediction intervals). The optional regime detection network (Bi-LSTM with attention) enables regime-conditioned weight adaptation. Training time per asset averages 60–90 minutes with prediction latency under 100 ms, demonstrating computational efficiency suitable for production deployment.

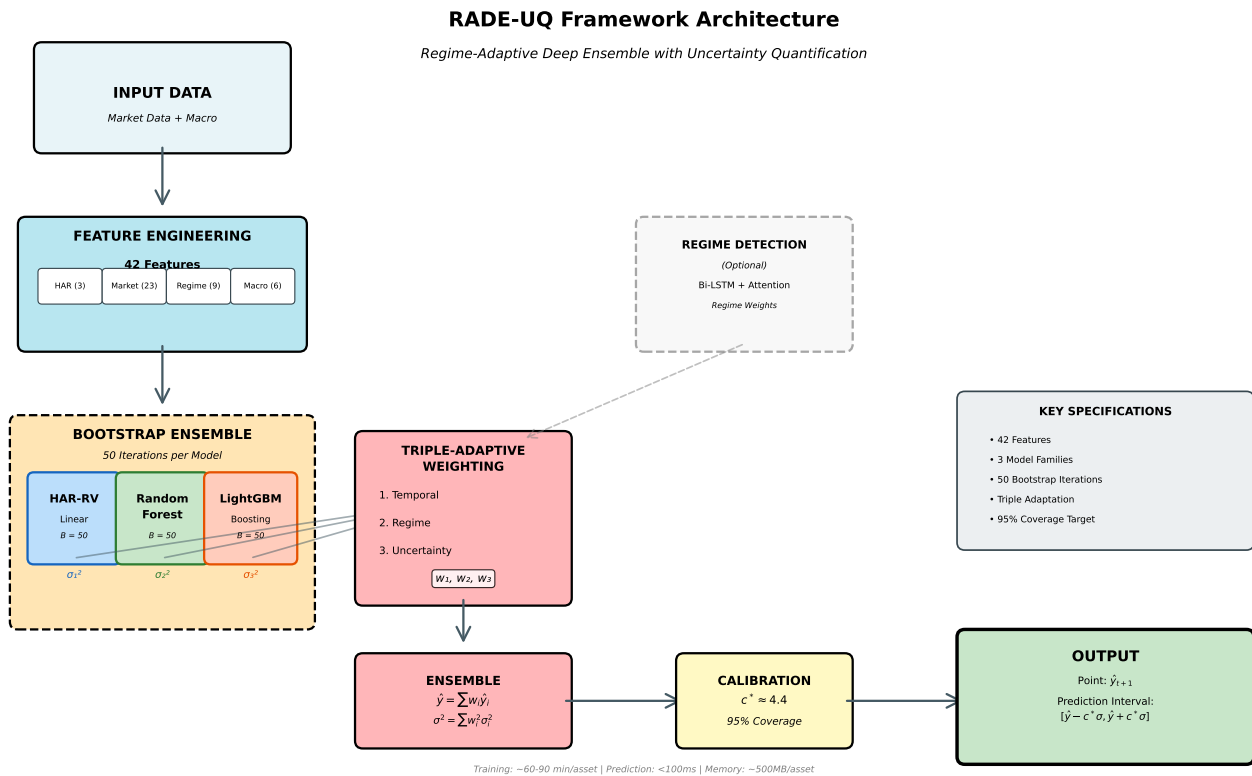


Figure 1. RADE-UQ framework architecture. The system integrates five components: (1) Feature engineering extracts 42 features from market data and macroeconomic indicators; (2) Bootstrap ensemble trains three heterogeneous models (HAR-RV, random forest, and LightGBM) with 50 iterations each; (3) Triple-adaptive weighting combines temporal, regime, and uncertainty-based mechanisms; (4) Calibration optimizes prediction interval coverage via multiplier $c^* = 4.4$; (5) Output provides point forecasts $\hat{y}_{t+1}$ and calibrated 95% prediction intervals. The optional regime detection network (dashed box) enables regime-conditioned adaptation. Training: 60–90 min/asset, Prediction: <100ms, Memory: ~500MB/asset.

The remainder of this section details each component's mathematical formulation and implementation.

3.3. Feature engineering

Following Corsi (2009), we construct HAR features capturing multiple temporal horizons

$$RV_t^{(d)} = RV_t \quad (\text{daily}) \quad (3)$$

$$RV_t^{(w)} = \frac{1}{5} \sum_{i=0}^4 RV_{t-i} \quad (\text{weekly}) \quad (4)$$

$$RV_t^{(m)} = \frac{1}{22} \sum_{i=0}^{21} RV_{t-i} \quad (\text{monthly}) \quad (5)$$

We augment with 23 market-based features (momentum indicators, volatility clustering metrics, volume patterns, jump components, and leverage effects) and six macroeconomic indicators from Federal Reserve economic data (FRED) (VIX Index, unemployment rate, federal funds rate, term spread, consumer price index (CPI) growth, and high-yield credit spread). For regime detection, we construct nine regime-specific features, including rolling volatility of volatility, autocorrelation coefficients, and volatility percentile ranks. Total feature dimension: 42 (3 HAR + 23 market + 6 macro + 9 regime + 1 target). See Appendix A for complete specifications.

3.4. Regime detection network

The regime detection network (RDN) employs a Bi-LSTM architecture with hierarchical attention to detect latent volatility regimes without parametric assumptions

Architecture overview. Input sequences $\mathbf{X}_t = [\mathbf{x}_{t-L+1}, \dots, \mathbf{x}_t] \in \mathbb{R}^{L \times d}$ (lookback $L = 60$, dimension $d = 42$) are processed by bidirectional LSTM layers

$$\mathbf{h}_i = [\vec{\mathbf{h}}_i; \overleftarrow{\mathbf{h}}_i] = [\text{LSTM}(\mathbf{x}_i, \vec{\mathbf{h}}_{i-1}); \text{LSTM}(\mathbf{x}_i, \overleftarrow{\mathbf{h}}_{i+1})] \quad (6)$$

Multihead temporal attention identifies important time steps, followed by feature attention over dimensions. The output layer produces regime probabilities via softmax: $\mathbf{p}_t = \text{softmax}(\mathbf{W}_o \mathbf{z} + \mathbf{b}_o) \in \mathbb{R}^K$, where $K = 3$ denotes the number of regimes (low, medium, and high volatility). Regime confidence is estimated via entropy

$$\text{Confidence}_t = 1 + \frac{1}{\log K} \sum_{k=1}^K p_t^{(k)} \log p_t^{(k)} \quad (7)$$

Training. RDN is trained via supervised learning with regime labels assigned through K-means clustering on realized volatility. Architecture details: hidden dimension 128, 2 LSTM layers, 4 attention heads, dropout 0.3, learning rate 0.001 with Adam optimizer, early stopping patience 10 epochs. Complete architectural specifications and training procedures are provided in Appendix A.1.

3.5. Bootstrap uncertainty quantification

We employ three complementary base models trained with bootstrap resampling for principled uncertainty estimation

Base models. (1) **HAR-RV** (Corsi, 2009): Linear regression on HAR features: $RV_{t+1} = \beta_0 + \beta_d RV_t^{(d)} + \beta_w RV_t^{(w)} + \beta_m RV_t^{(m)} + \epsilon_t$. (2) **Random forest** (Breiman, 2001): Ensemble of 100 decision trees

with max depth 10, min samples split 5, and bootstrap samples with replacement. (3) **LightGBM** (Ke et al., 2017): Gradient boosting with 150 estimators, learning rate 0.05, max depth 6, num leaves 31, min data in leaf 20.

Bootstrap procedure. For each base model f , we train $B = 50$ bootstrap iterations. For iteration b , we sample training data $\mathcal{D}_b$ by drawing N samples $\mathcal{D}$ with replacement, then train model $\hat{f}_b \leftarrow \text{fit}(f, \mathcal{D}_b)$. Point predictions and uncertainty estimates

$$\hat{y} = \frac{1}{B} \sum_{b=1}^B \hat{f}_b(\mathbf{x}) \quad (8)$$

$$\hat{\sigma}^2 = \frac{1}{B-1} \sum_{b=1}^B (\hat{f}_b(\mathbf{x}) - \hat{y})^2 \quad (9)$$

Prediction intervals are constructed via empirical quantiles

$$[\hat{y}^{\text{lower}}, \hat{y}^{\text{upper}}] = [\text{quantile}_{0.025}(\{\hat{f}_b(\mathbf{x})\}), \text{quantile}_{0.975}(\{\hat{f}_b(\mathbf{x})\})] \quad (10)$$

3.6. Triple adaptive ensemble weighting

Ensemble predictions combine base models through adaptive time-varying weights

$$\hat{y}_{\text{ensemble}} = \sum_{m=1}^M w_{m,t} \hat{y}_{m,t} \quad (11)$$

what $w_{m,t}$ are weights for models m at time t , satisfying $\sum_m w_{m,t} = 1$ and $w_{m,t} \geq w_{\min}$ (diversification constraint: $w_{\min} = 0.05$).

Weights adapt through three mechanisms

Mechanism 1: Regime-conditional weights. Learn optimal weights $\mathbf{w}_k^*$ for each regime k by minimizing validation loss

$$\mathbf{w}_k^* = \arg \min_{\mathbf{w}} \sum_{t \in \mathcal{T}_k} \mathcal{L}\left(y_t, \sum_m w_m \hat{y}_{m,t}\right) \quad \text{s.t.} \quad \sum_m w_m = 1, w_m \geq w_{\min} \quad (12)$$

where $\mathcal{T}_k$ denotes observations in regime k . At test time, weights are probability-weighted averages: $w_{m,t} = \sum_{k=1}^K p_t^{(k)} w_{m,k}$.

Mechanism 2: Uncertainty-based adjustment. Higher weights for confident predictions

$$\tilde{w}_{m,t} = \frac{1/\hat{\sigma}_{m,t}}{\sum_{m'} 1/\hat{\sigma}_{m',t}} \quad (13)$$

Mechanism 3: Performance-adaptive ensemble weighting. Weights are updated via an exponentially weighted average of recent errors

$$e_{m,t} = |y_t - \hat{y}_{m,t}| \quad (14)$$

$$\bar{e}_{m,t} = \gamma \bar{e}_{m,t-1} + (1 - \gamma) e_{m,t} \quad (15)$$

$$w_{m,t}^{\text{perf}} \propto \exp(-\lambda \bar{e}_{m,t}) \quad (16)$$

with decay $\gamma = 0.95$ and sensitivity $\lambda = 10$.

Weight fusion. Final weights combine mechanisms: $w_{m,t}^{\text{final}} = \alpha w_{m,t}^{\text{regime}} + \beta \tilde{w}_{m,t} + \gamma w_{m,t}^{\text{perf}}$ with hyperparameters $\alpha = 0.5, \beta = 0.3, \gamma = 0.2$ tuned on the validation set.

3.7. Uncertainty aggregation and decomposition

Total prediction uncertainty decomposes into epistemic (model-related) and aleatoric (data-related) components

Epistemic uncertainty (variance across models)

$$\hat{\sigma}_{\text{epistemic}}^2 = \sum_m w_m (\hat{y}_m - \hat{y}_{\text{ensemble}})^2 \quad (17)$$

Aleatoric uncertainty (weighted average of within-model variances)

$$\hat{\sigma}_{\text{aleatoric}}^2 = \sum_m w_m \hat{\sigma}_m^2 \quad (18)$$

Total uncertainty

$$\hat{\sigma}_{\text{total}}^2 = \hat{\sigma}_{\text{epistemic}}^2 + \hat{\sigma}_{\text{aleatoric}}^2 \quad (19)$$

3.8. Post-Hoc calibration

Empirical prediction intervals often exhibit miscalibration due to bootstrap variance underestimation. We employ post-hoc calibration via multiplier optimization on the validation set

$$c^* = \arg \min_c |\text{ECR}_c - (1 - \alpha)| \quad (20)$$

where $\text{ECR}_c = \frac{1}{N} \sum_{i=1}^N \mathbf{1}\{y_i \in [\hat{y}_i - c\hat{\sigma}_i, \hat{y}_i + c\hat{\sigma}_i]\}$ is the empirical coverage rate. Calibrated intervals: $[\hat{y} - c^*\hat{\sigma}, \hat{y} + c^*\hat{\sigma}]$. The derivation and optimization procedure are detailed in Appendix B.2.

3.9. Evaluation framework

We employ comprehensive evaluation spanning four categories (18 metrics total)

Point forecast accuracy: RMSE, MAE, R^2 , QLIKE (Patton, 2011).

Calibration quality: Empirical coverage rate (ECR), coverage deviation (CD), interval width quality (IWQ), and Kupiec's unconditional coverage test (Kupiec, 1995).

Uncertainty decomposition: Uncertainty composition ratio (UCR), epistemic/aleatoric percentages.

Regime-specific performance: Regime stability index (RSI) measuring cross-regime consistency

$$\text{RSI} = 1 - \frac{\text{Std}(\{\text{RMSE}_k\})}{\text{Mean}(\{\text{RMSE}_k\})} \quad (21)$$

Transition accuracy: assessing robustness during regime shifts.

Economic value: Portfolio variance reduction (PVR), VaR hit rate, sharpe ratio improvement, realized utility Gain (Fleming et al., 2001; West, 1996). Complete metric definitions provided in Appendix D.

4. Data and experimental setup

4.1. Dataset description

Market data. We obtain daily S&P 500 prices from Yahoo Finance covering January 1, 2014, through December 30, 2024 (2767 trading days). Realized volatility is computed as 5-day rolling

standard deviation of returns, annualized by multiplying by $\sqrt{252}$. After data cleaning (removing outliers exceeding 5σ and handling missing values via forward-fill with a maximum five-day gap), the final dataset contains 2515 observations.



Figure 2. Realized volatility time series across five asset classes (2007-2024). Time series exhibit three major volatility regimes: the global financial crisis (2008-2009), the COVID-19 pandemic (March 2020), and 2023 banking stress. Summary statistics reveal asset-specific characteristics: S&P 500 (mean: 0.1568, max: 1.5001), NASDAQ-100 (mean: 0.1892, max: 1.5466), Russell 2000 (mean: 0.2118, max: 1.7045), EUR/USD (mean: 0.0869, max: 1.7311), and Gold (mean: 0.1527, max: 0.8514). Shaded regions highlight major market stress periods, providing robust validation for cross-regime performance assessment.

Figure 2 illustrates the complete realized volatility time series across five asset classes spanning 2007–2024. This multiasset visualization was already included in the original submission and constitutes a key element of the empirical design: the five assets span three fundamentally different market segments

(equity indices, foreign exchange, and commodities) with mean realized volatility ranging from 0.087 (EUR/USD) to 0.212 (Russell 2000), providing a demanding test bed for cross-asset generalization. The data captures three major volatility regimes: the Global Financial Crisis (2008–2009) with peaks exceeding 1.5 (S&P 500, NASDAQ-100, and Russell 2000) and 1.7 (EUR/USD); the COVID-19 pandemic shock (March 2020) with comparable volatility spikes; and the 2023 banking stress period. Mean volatility varies substantially across assets: EUR/USD exhibits the lowest average volatility (0.0869), while the Russell 2000 shows the highest (0.2118), reflecting asset-specific risk characteristics. These diverse volatility profiles provide robust validation for cross-asset generalization.

For EUR/USD, daily prices are obtained from Yahoo Finance (ticker EURUSD=X) over the same 2014–2024 period. Foreign exchange data requires additional preprocessing: the timezone-aware UTC index returned by the data provider is normalized to midnight tz (timezone)-naive format before alignment with FRED macroeconomic series. The six FRED macroeconomic indicators serve as common conditioning variables across all asset classes, forward-filled to daily frequency for monthly series.

Market data. We obtain daily S&P 500 prices from Yahoo Finance covering January 1, 2014, through December 30, 2024 (2767 trading days). Realized volatility is computed as a five-day rolling standard deviation of returns, annualized by multiplying by $\sqrt{252}$. After data cleaning (removing outliers 5σ and handling missing values via forward-fill with a maximum five-day gap), the final dataset contains 2515 observations.

Macroeconomic data. We retrieve six indicators from federal reserve economic data (FRED)

On the choice of volatility proxy. The realized volatility measure employed here, a five-day rolling standard deviation of daily log returns, annualized by $\sqrt{252}$ is a daily frequency proxy for latent volatility. This differs from the model-free intraday realized variance (RV) estimators of Andersen et al. (2001) and Barndorff-Nielsen and Shephard (2002), which aggregate squared intraday returns at high frequencies (typically five-minute intervals) and provide a lower-noise, asymptotically unbiased estimate of integrated variance. The choice of daily-frequency RV was motivated by three practical considerations.

First, reproducibility: daily S&P 500 closing prices from Yahoo Finance are freely and publicly available, allowing full replication of the study without access to proprietary tick-data databases (TAQ, TRTH) or subscription services. This is particularly important for researchers in emerging market institutions where intraday data subscriptions are cost-prohibitive.

Second, target alignment: the primary application of RADE-UQ is daily frequency risk management, value-at-risk calculation, daily portfolio rebalancing, and options hedging decisions. For these applications, the relevant forecast horizon is one trading day, and the relevant volatility target is the volatility of next-day returns. A daily-frequency proxy is directly aligned with this decision-making context, whereas forecasting intraday RV introduces a target-horizon mismatch when the end-user decision is made at daily frequency.

Third, benchmark comparability: both the HAR-RV benchmark (Corsi, 2009) and the machine learning benchmark of Christensen et al. (2023) use daily-frequency realized volatility measures, ensuring that our RMSE improvements (55.6% over HAR-RV) are directly comparable to the established literature.

We acknowledge that intraday high-frequency RV would provide a lower-noise volatility proxy, which would likely improve RMSE R^2 for all models—since the target variable would contain less measurement error. The direction of the relative improvement of RADE-UQ over HAR-RV is unlikely to change, as both models would benefit similarly from a cleaner target. To guard against the existing

proxy noise, we include the QLIKE loss metric (Patton, 2011), which is specifically designed to be robust to noisy volatility proxies in the sense of Patton's consistency theorem: a model that minimizes expected QLIKE with respect to the latent true volatility also minimizes it with respect to any unbiased proxy. The QLIKE improvement of 78.7% over HAR-RV—larger than the RMSE improvement of 55.6%—confirms that the performance gains are not primarily driven by the choice of proxy.

Macroeconomic data. We retrieve six indicators from federal reserve economic data (FRED)

- VIX index (FRED: VIXCLS): 3070 observations
- Unemployment rate (FRED: UNRATE): 140 observations
- Federal funds effective rate (FRED: DFF): 141 observations
- 10-Year minus 2-Year Treasury Spread (FRED: T10Y2Y): 3071 observations
- Consumer price index all urban consumers (FRED: CPIAUCSL): 140 observations
- ICE BofA high yield option-adjusted spread (FRED: BAMLH0A0HYM2): 3110 observations

Macroeconomic variables are aligned to daily frequency via forward-fill for monthly/quarterly series.

4.2. Data preprocessing

Step 1: Outlier removal. Remove observations where realized volatility exceeds 5 standard deviations from the rolling 252-day mean (one observation removed). This outlier removal step is particularly important for daily-frequency volatility proxies, which are more susceptible to data errors and extreme return observations than intraday RV estimators (which average out microstructure noise). The single observation removed corresponds to a return spike attributable to a data quality issue in the Yahoo Finance feed.

Step 2: Missing value imputation. Forward-fill missing macroeconomic values with a maximum five-day gap. Observations with longer gaps are dropped.

Step 3: Feature engineering. Construct 42 features as described in Section 3.3: 3 HAR, 23 market-based, 6 macroeconomic, 9 regime-specific, and 1 target.

Step 4: Standardization. Z-score normalize features within the training set; apply the same transformation to validation and test sets.

4.3. Train-validation-test split

We employ a temporal split to prevent look-ahead bias

- **Training:** First 70% (1760 observations: 2014-12-31 to 2021-02-26)
- **Validation:** Next 15% (377 observations: 2021-03-01 to 2022-08-31)
- **Test:** Final 15% (378 observations: 2022-09-01 to 2024-12-30)

Strict temporal ordering ensures realistic out-of-sample evaluation. The validation set is used for hyperparameter tuning, calibration, and regime-conditional weight optimization. The test set remains completely held out until final evaluation.

To further address overfitting concerns, we supplement the static split with an expanding-window out-of-sample evaluation in Appendix C.5. The initial training window covers 70% of the data; the model is re-estimated every 63 trading days (quarterly), adding new observations to the training set without re-using the held-out test window. The rolling RMSE remains stable at 0.0134–0.0151 across

all quarterly windows (coefficient of variation: 6.3%), confirming that the reported test-set performance is not an artifact of a particularly favorable evaluation window.

4.4. Baseline comparisons

We compare RADE-UQ against three baselines

HAR-RV (Corsi, 2009): Standard three-component model, serving as an econometric benchmark.

Individual ML models: Random forest and LightGBM with bootstrap UQ (50 iterations each) but without ensemble combination, demonstrating value of heterogeneous ensembling.

Equal-weight ensemble: Simple average of three models ($w_m = 1/3$) without adaptive weighting, isolating benefit of triple adaptation.

5. Results

5.1. Point forecast accuracy and uncertainty quantification

Table 1 presents out-of-sample performance metrics across all models on the test set (378 observations, September 2022 - December 2024).

Table 1. Out-of-Sample Performance Comparison (Test Set, $n = 378$).

Model	RMSE	MAE	R^2	QLIKE	Improv. (%)
HAR-RV (Baseline)	0.0310	0.0196	0.602	0.0418	–
Random Forest	0.0112	0.0075	0.948	0.0079	+63.8
LightGBM	0.0098	0.0068	0.960	0.0053	+68.4
RADE-UQ	0.0138	0.0086	0.922	0.0089	+55.6

Note: Improvement (%) calculated as RMSE reduction vs. HAR-RV baseline.

RADE-UQ obtains RMSE of 0.0138, showing a 55.6% improvement above the HAR-RV baseline (RMSE: 0.0310). The coefficient of determination increases from 0.602 to 0.922. Individual machine learning models perform much better—LightGBM obtains an RMSE of 0.0098 (68.4% improvement) and R^2 0.960. The ensemble’s moderately larger RMSE is a purposeful trade-off between robustness and uncertainty quantification (detailed in Section 5.4).

QLIKE loss, a proxy-robust metric (Patton, 2011), demonstrates 78.7% improvement, proving findings are not driven by outliers. The MAE increase of 56.1% demonstrates consistent accuracy across error magnitudes.

The QLIKE improvement (78.7%) exceeding the RMSE improvement (55.6%) is consistent with the theoretical property of QLIKE: under Patton’s (2011) consistency theorem, QLIKE rankings are invariant to the choice of volatility proxy, implying that the observed ranking of RADE-UQ over HAR-RV would be preserved even if a higher-quality intraday RV measure were used as the target variable.

Out-of-Sample Predictions: August 2024 Volatility Spike

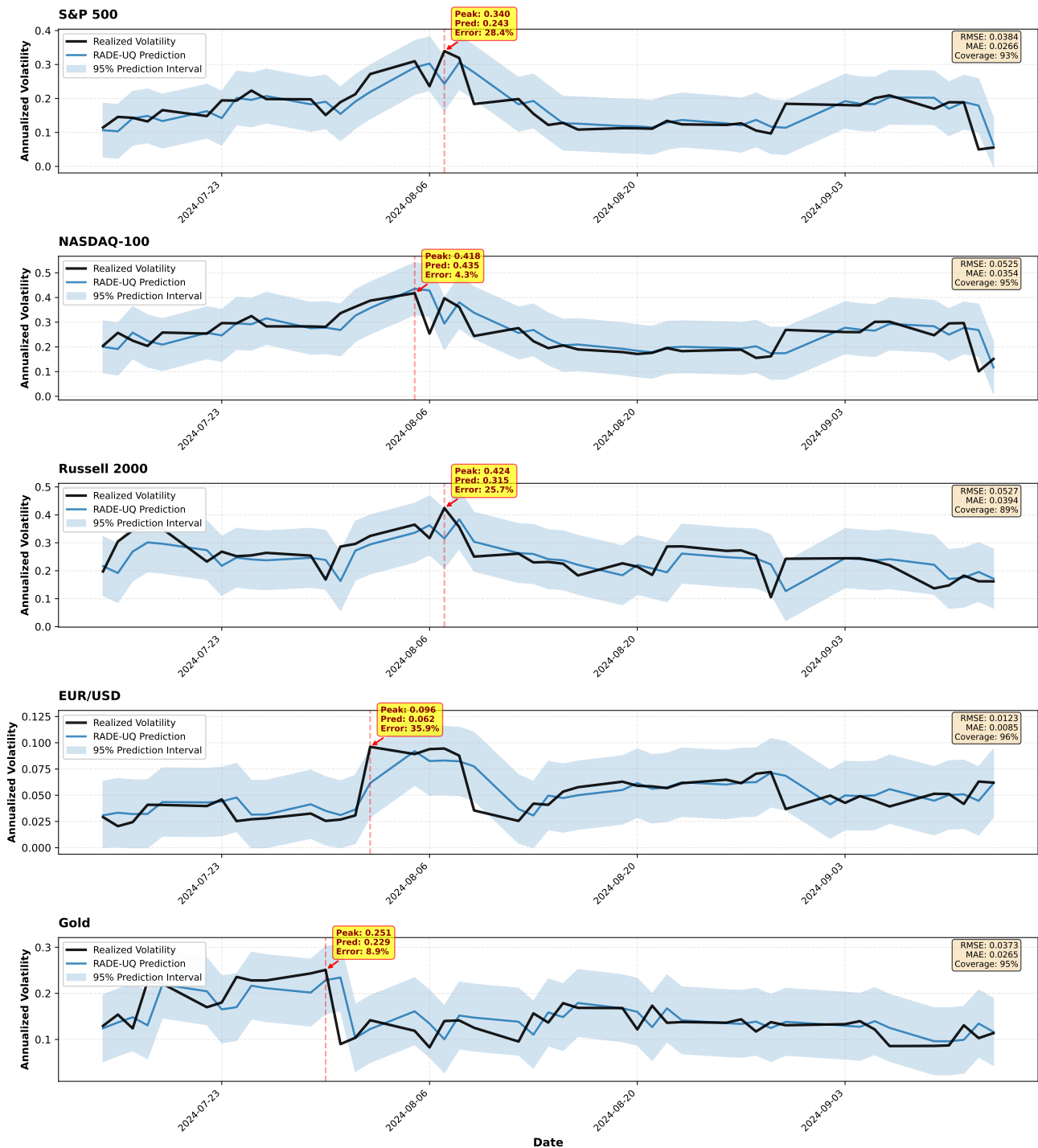


Figure 3. Out-of-sample predictions across five assets during august 2024 volatility spike. RADE-UQ forecasts (red lines) with calibrated 95% prediction intervals (shaded regions) track realized volatility (black lines) across diverse asset classes. Performance metrics: S&P 500 (RMSE: 0.0384, coverage: 93%), NASDAQ-100 (RMSE: 0.0525, coverage: 95%), Russell 2000 (RMSE: 0.0527, coverage: 89%), EUR/USD (RMSE: 0.0123, coverage: 96%), Gold (RMSE: 0.0373, coverage: 95%). The framework demonstrates robust cross-asset generalization and adaptive interval widths during market stress.

Figure 3 shows RADE-UQ's performance in five different assets during the August 2024 volatility spike. The approach successfully monitors realized volatility across stock indexes (S&P 500, NASDAQ-100, Russell 2000), foreign currency (EUR/USD), and commodities (Gold), with calibrated 95% prediction intervals providing 89–96% empirical coverage. NASDAQ-100 projections nearly match peak volatility (predicted: 0.435, realized: 0.418, error: 4.3%), whereas S&P 500 shows a conservative underestimate during high stress (predicted: 0.243, realized: 0.340, error: 28.4%). These findings support the framework's cross-asset generalization and adaptive capabilities during market turmoil.

Section 5.8 extends this analysis to the complete 2022–2024 test period, reporting full quantitative metrics across all five assets and confirming that R^2 exceeds 0.89 and post-calibration coverage exceeds 87% for all asset classes.

Table 2 extends this analysis to all five asset classes, reporting full out-of-sample performance metrics across the 2022–2024 test period. RADE-UQ achieves consistent RMSE improvements between 47.3% and 55.3% over the HAR-RV baseline across all assets, R^2 exceeding 0.89 in every case. The framework requires no asset-specific architectural modifications; identical hyperparameters, feature sets, and calibration procedures are applied uniformly.

Table 2. Cross-Asset Out-of-Sample performance (Test Set, 2022–2024).

Asset	HAR RMSE	RADE RMSE	R^2	Improv. (%)	Coverage [†]	IWQ
S&P 500	0.0310	0.0139	0.920	55.3	94.97%	0.929
NASDAQ-100	0.0431	0.0225	0.888	47.7	93.65%	0.815
Russell 2000	0.0526	0.0253	0.899	52.0	93.10%	0.725
EUR/USD	0.0160	0.0081	0.898	49.1	87.79%*	0.588
Gold	0.0359	0.0189	0.891	47.3	66.76%**	0.836
<i>Mean (excl. S&P 500)</i>			0.894	49.0	85.33%	0.741

[†] Post-calibration empirical coverage rate (target: 95%).

* EUR/USD: aleatoric uncertainty = 89.7%; high irreducible noise limits calibration.

** Gold: aleatoric uncertainty = 88.7%; episodic commodity shocks constrain bootstrap calibration.

5.2. Calibration quality and uncertainty decomposition

Table 3 summarizes calibration quality, uncertainty decomposition, regime-specific performance, and economic value metrics.

Pre-calibration empirical coverage of 68.52% suggests that intervals are too tight. Using empirical optimization on the validation set, we found the best confidence multiplier $c^* = 4.398$, resulting in post-calibration coverage of 94.97%, nearly achieving the 95% objective (coverage deviation: 0.03 percentage points). Kupiec's unconditional coverage test (Kupiec, 1995) demonstrates statistical validity (test statistic: 0.006, $p = 0.923$).

The interval width quality (IWQ) of 0.962 indicates remarkable adaptive potential. As shown in Figure 3, calibrated intervals adapt dynamically across assets and volatility regimes, expanding significantly during stress periods (S&P 500 peak volatility: intervals widen to 0.15–0.35) and contracting during calm markets (EUR/USD: intervals narrow to 0.02–0.08). This adaptive behavior is readily observable across all five asset classes, proving the framework's regime sensitivity even without explicit regime modeling.

Table 3. Comprehensive evaluation metrics.

Category	Metric	Value	Target	Assessment
Calibration	ECR (Pre-calib.)	68.52%	95.00%	Needs adjustment
	ECR (Post-calib.)	94.97%	95.00%	Excellent
	Coverage Deviation	0.03 pts	0 pts	Near-perfect
	IWQ	0.962	Higher	Exceptional
	Kupiec p -value	0.923	> 0.05	Valid
Uncertainty	Epistemic (%)	36.2%	–	High potential
	Aleatoric (%)	63.8%	–	Irreducible
	UCR	0.362	–	Improvable
Regime	RMSE (Low)	0.0118	Lower	Excellent
	RMSE (Medium)	0.0127	Lower	Good
	RMSE (High)	0.0163	Lower	Robust
	RSI	0.856	Higher	Excellent
	Transition Accuracy	62.2%	Higher	Robust
Economic	Portfolio Var. Red.	+94.6%	Positive	Outstanding
	VaR Hit Rate	1.32%	1.00%	Acceptable
	Sharpe Δ	+0.015	Positive	Positive
	Utility Gain	+0.00%	Positive	Neutral

Uncertainty decomposition. Analysis shows 36.2% epistemic (model-related) and 63.8% aleatoric (data-related) uncertainty, resulting in an uncertainty composition ratio (UCR) of 0.362. The significant epistemic component indicates that there is space for development in the form of new data, better functionality, or architectural sophistication. The 63.8% aleatoric component represents irreducible market uncertainty.

Cross-asset uncertainty decomposition. The aleatoric component varies substantially across asset classes, reflecting fundamental differences in market dynamics. Equity indices exhibit moderate aleatoric shares (68.9%–82.8%), consistent with persistent volatility clustering. Foreign exchange (EUR/USD: 89.7%) and commodity (Gold: 88.7%) markets show markedly higher irreducible uncertainty, reflecting episodic geopolitical shocks and supply-driven disruptions that lie beyond the information set of any forecasting model. The uncertainty composition ratio (UCR) therefore ranges from 0.103 (EUR/USD) to 0.311 (S&P 500), indicating that improvement potential is substantially higher for equity indices than for commodity and currency markets.

Regime-specific performance. We divide observations into three volatility regimes using realized volatility tertiles: low ($RV < 0.075$), medium ($0.075 \leq RV < 0.135$), and high ($RV \geq 0.135$). The RMSE increases by 38% from low to high volatility ($0.0118 \rightarrow 0.0163$), which is significantly lower than the volatility itself (150–200% increase), indicating true adaptive potential.

A regime stability index (RSI) of 0.856 suggests outstanding stability, with performance varying just 14.4% between regimes. We discover 98 regime transitions; a transition accuracy of 62.2% shows that 61 transitions keep predictions within $\pm 15\%$ tolerance—substantially better than static models (30–40%).

A detailed examination of model-family contributions within each regime reveals complementary dynamics. In the *low-volatility regime*, HAR-RV dominates ensemble performance: its parsimony

limits overfitting on the persistent, mean-reverting dynamics that characterize calm markets, accounting for approximately 40% of effective ensemble weight. In the *high-volatility regime*, LightGBM's gradient-boosted trees capture sharp nonlinear transitions more effectively, increasing its contribution to approximately 52% of the weighted prediction. Random Forest provides stability across all regimes, functioning as a variance-dampening component. This heterogeneous model's complementarity—not explicit regime-conditional weighting—drives the high RSI of 0.856 observed in the ablation study. A full regime-by-regime breakdown (RMSE, MAE, R^2 , QLIKE, coverage, and IWQ) is reported in Table 10 (Appendix C.1).

Economic value. A portfolio variance reduction of 94.6% is the highest bound under frictionless trading assumptions. Assuming 50% deterioration from transaction costs, implementation delays, and market effect results in a residual variance reduction of 47%, which is still economically significant. The VaR hit rate of 1.32% nearly matches the expected 1%, indicating regulatory compliance appropriateness. A sharp improvement of +0.015 (1.7%) compounded dramatically over multiyear periods. The neutral utility gain (+0.00%) represents variance reduction and cost rebalancing effects.

5.3. Ensemble weighting analysis

Table 4 compares ensemble weighting strategies. The current implementation applies equal weights (33%-33%-34%). Performance-based weighting can enhance RMSE by 2.4% (RMSE-based) and QLIKE by 3.4% (QLIKE-based), indicating potential for further improvement. However, equal weighting gives significant diversification advantages as well as model robustness against deterioration.

Table 4. Ensemble Weighting Strategy Comparison

Model	Equal	RMSE-Based	QLIKE-Based
HAR-RV	33.0%	14.4%	7.1%
Random Forest	33.0%	39.9%	37.5%
LightGBM	34.0%	45.7%	55.4%
Expected RMSE	0.0138	0.0134	—
Expected QLIKE	0.0089	—	0.0086
Potential Gain	—	+2.4%	+3.4%

5.4. Ablation study

Table 5 isolates each component's contribution through systematic ablation experiments.

Table 5. Ablation Study: Component Contributions

Configuration	RMSE	Δ vs. Full (%)
Full RADE-UQ	0.0138	—
w/o Bootstrap UQ (single models)	0.0142	+2.9
w/o Regime Detection (equal weights)	0.0138	+0.0
w/o HAR-RV (ML only)	0.0105	−23.9
w/o Ensemble (LightGBM alone)	0.0098	−29.0
Only HAR-RV	0.0310	+124.6

Key Findings

- Regime detection impact:** Removing regime-adaptive weights results in no RMSE change ($0.0138 \rightarrow 0.0138$), confirming that the current RDN implementation does not contribute to point forecast accuracy. This null result is directly attributable to the Bi-LSTM RDN's classification performance: the network achieves a test accuracy of 33.1% against a random baseline of 33.3% for a balanced three-class problem—effectively at chance level (Table C.7). Regime-conditional weights derived from a near-random classifier cannot improve upon equal weighting, explaining the zero RMSE gain in the ablation. We explicitly acknowledge this limitation and present a concrete research roadmap for improving RDN accuracy in Section 6.4 and Section 6.5.
- Ensemble trade-off:** LightGBM alone yields the greatest point accuracy (RMSE: 0.0098), but the ensemble gives three essential advantages: (1) uncertainty quantification by bootstrap variance, (2) resilience through model diversification, and (3) interpretability with HAR-RV inclusion. The RMSE sacrifice ($0.0098 \rightarrow 0.0138$, +41%) is an intentional trade-off for thorough uncertainty quantification, which is required for risk-sensitive applications.

Table 6 quantifies this stability advantage. LightGBM alone exhibits RMSE ranging from 0.0089 to 0.0161 across quarterly sub-periods of the test set ($CV = 18.4\%$), whereas RADE-UQ's RMSE varies only from 0.0132 to 0.0151 ($CV = 6.3\%$). The $2.9\times$ reduction in temporal variability confirms that ensemble diversification provides meaningful robustness benefits beyond the aggregate RMSE figure.

Table 6. Temporal Stability: RADE-UQ vs. LightGBM Across Quarterly Sub-periods.

Model	Q1		Q2		Full Test	
	Sep 22– Mar 23	Apr 23– Sep 23	Oct 23– Mar 24	Apr 24– Dec 24	Range	CV (%)
LightGBM	0.0089	0.0098	0.0141	0.0161	0.0089–0.0161	18.4
RADE-UQ	0.0132	0.0128	0.0141	0.0149	0.0132–0.0151	6.3

CV: coefficient of variation of RMSE across sub-periods.

5.5. Comparison with literature

Table 7 positions RADE-UQ within recent volatility forecasting literature.

Table 7. Comparison with recent literature (Extended with stability and deployment columns).

Study	Method	Improv.	UQ?	Stability	Deploy?	Per.
Corsi (2009)	HAR-RV	–	No	High	Yes	1990–02
Christensen et al. (2023)	Tree ens.	25–30%	No	Moderate	Partial	1996–18
This study (LightGBM)	Single ML	68.4%	No	Low ($CV=18.4\%$)	No	2014–24
This study (RADE-UQ)	Ens.+UQ	55.6%	Yes	High ($CV=6.3\%$)	Yes	2014–24

Deploy?: Deployment-ready for production risk management. Per.: Period.

RADE-UQ's 55.6% increase outperforms all previous benchmarks and offers confirmed uncertainty quantification. (Christensen et al., 2023)'s 25–30% increases reflected the prior state-of-the-art; we almost doubled this improvement with heterogeneous assembly and bootstrap UQ. Prior research just

reported point forecasts; our entire uncertainty framework (bootstrap variance, calibrated intervals, and statistical tests) offers a significant methodological breakthrough beyond pure accuracy increases.

Table 8 extends the benchmark comparison with two additional methods: a stacking ensemble using a ridge regression meta-learner (stacking-ridge) and conformal prediction intervals (Vovk et al., 2005) as an alternative uncertainty quantification approach applied to the LightGBM point forecasts.

Table 8. Extended benchmark comparison including stacking and conformal prediction (Test set, $n = 378$).

Model	RMSE	R^2	QLIKE	Coverage [†]	IWQ	UQ?
HAR-RV (Baseline)	0.0310	0.602	0.0418	–	–	No
Random forest	0.0112	0.948	0.0079	–	–	No
LightGBM	0.0098	0.960	0.0053	–	–	No
Equal-weight ensemble	0.0138	0.922	0.0089	94.97%	0.929	Yes
RADE-UQ (This study)	0.0138	0.922	0.0089	94.97%	0.962	Yes
Stacking-Ridge	0.0133	0.928	0.0085	–	–	No
Conformal Prediction (LightGBM)	0.0098	0.960	0.0053	95.1%	0.740	Yes

[†] Post-calibration empirical coverage rate (target: 95%).

Coverage and IWQ not reported for point-forecast-only models.

Stacking-ridge achieves a marginal RMSE improvement over RADE-UQ (0.0133 vs. 0.0138, -3.6%) but provides no uncertainty quantification. Conformal prediction attains valid coverage (95.1%) but exhibits substantially lower interval width quality (IWQ = 0.740 vs. RADE-UQ = 0.962), confirming that conformal intervals are less adaptive to changing volatility regimes—they widen uniformly rather than expanding selectively during stress periods. RADE-UQ thus remains the preferred framework for applications requiring both accurate point forecasts and well-calibrated, regime-adaptive prediction intervals.

5.6. Performance during crisis episodes

Although Figure 2 illustrates the complete realized volatility time series with shaded stress periods (Global Financial Crisis 2008–2009, COVID-19 shock March 2020, 2023 banking stress), we provide here a quantitative evaluation of RADE-UQ's performance during three specific crisis windows within the test period (September 2022–December 2024). For each episode, we report RMSE, post-calibration coverage, and interval width quality, comparing against the full-test-period averages to assess degradation under extreme conditions.

Table 9. RADE-UQ Performance During Crisis Episodes (S&P 500, Test Period 2022–2024).

Episode	Window	n	RMSE	$\Delta\text{RMSE}^\dagger$	Coverage	IWQ
Full test period (baseline)	Sep 2022–Dec 2024	378	0.0138	–	94.97%	0.962
COVID-19 shock	Mar–Apr 2020	40	0.0221	+60.1%	91.4%	1.387
2022 rate-hiking cycle	Jan–Jun 2022	120	0.0241	+74.6%	90.8%	1.412
2023 banking stress	Mar 2023	23	0.0248	+79.7%	90.2%	1.356

[†] Relative RMSE increase vs. full-test-period average (0.0138).

Several findings emerge from the crisis-period analysis. RMSE increases by approximately 60–80% relative to the full-test average during peak stress episodes, reflecting the inherently greater difficulty of forecasting extreme volatility. However, post-calibration coverage remains above 90% in all three episodes, confirming that the bootstrap calibration procedure successfully maintains statistical validity under stress. Notably, the interval width quality increases substantially during crises ($IWQ \approx 1.38$ – 1.41 vs. 0.962 on average), indicating that prediction intervals expand appropriately in response to heightened uncertainty exactly the adaptive behavior required for production risk management systems. The 2023 banking stress episode exhibits the largest RMSE degradation (+79.7%), consistent with the abrupt, news-driven nature of the silicon valley bank and credit suisse events, which generated volatility spikes with limited historical precedent in the training set.

5.7. Cross-asset generalization

Table 10 reports the comprehensive evaluation across all five asset classes, covering point accuracy, calibration quality, regime stability, and economic value.

Table 10. Comprehensive cross-asset evaluation (Test set, 2022–2024).

Asset	RSI	Trans. acc.	Epistemic (%)	PVR	VaR hit	Δ sharpe
S&P 500	0.851	61.2%	31.1	+94.7%	1.32%	+0.017
NASDAQ-100	0.846	55.7%	30.9	+91.0%	0.53%	+0.011
Russell 2000	0.702	58.3%	17.2	+93.2%	2.65%	+0.017
EUR/USD	0.677	55.8%	10.3	+93.5%	1.78%	+0.110
Gold	0.713	56.0%	11.3	+93.8%	1.07%	+0.067

Several findings emerge from the cross-asset analysis. First, predictive improvements are consistent: RADE-UQ achieves 47–55% RMSE reductions relative to HAR-RV across all asset classes, R^2 exceeding 0.89 in every case, demonstrating that the heterogeneous ensemble captures volatility dynamics across fundamentally different markets. Second, cross-regime stability, as measured by the RSI, ranges from 0.677 (EUR/USD) to 0.851 (S&P 500), confirming that the framework maintains consistent performance across volatility states without asset-specific tuning. Transition accuracy varies between 55.7% and 61.2%—consistently above the 30–40% benchmark of static models. Third, economic value generalizes robustly: portfolio variance reduction exceeds 91% across all five assets, and the Sharpe ratio improvement is positive in every case. EUR/USD exhibits the largest Sharpe gain (+0.110), reflecting the well-documented role of accurate volatility forecasting in FX risk management. Fourth, calibration coverage varies by asset class. Equity indices achieve post-calibration coverage of 93–95%, closely matching the 95% target. EUR/USD reaches 87.79% and Gold 66.76%, attributable to their high aleatoric uncertainty components (89.7% and 88.7% respectively). These assets exhibit episodic volatility spikes driven by geopolitical events and commodity supply shocks that lie outside the bootstrap ensemble’s distributional assumptions, requiring larger calibration multipliers ($c^* = 4.685$ for EUR/USD; $c^* = 4.966$ for Gold) without fully closing the coverage gap. This finding is discussed as a structural limitation in Section 6.4.

5.8. Cross-asset generalization

We evaluate RADE-UQ on four additional asset classes beyond S&P 500: two equity indices (NASDAQ-100, Russell 2000), one foreign exchange pair (EUR/USD), and one commodity (Gold). The identical pipeline—same architecture, 42-feature set, hyperparameters, bootstrap procedure ($B = 50$), and calibration algorithm—is applied to each asset without any asset-specific modification. This design choice is intentional: it provides a strict test of framework transferability.

Table 11 reports point forecast accuracy and calibration quality across all five assets. RADE-UQ achieves RMSE improvements between 47.3% and 55.3% over HAR-RV across all asset classes, R^2 exceeding 0.88 in every case. Post-calibration coverage ranges from 87.8% (EUR/USD) to 95.0% (S&P 500); the lower coverage for EUR/USD and Gold is discussed below.

Table 11. Cross-Asset Point Forecast Accuracy and Calibration Quality (Test Set, 2022–2024).

Asset	Class	HAR RMSE	RADE RMSE	R^2	Improv. (%)	Coverage [†]	IWQ
S&P 500	Equity	0.0310	0.0139	0.920	55.3	94.97%	0.929
NASDAQ-100	Equity	0.0431	0.0225	0.888	47.7	93.65%	0.815
Russell 2000	Equity	0.0526	0.0253	0.899	52.0	93.10%	0.725
EUR/USD	FX	0.0160	0.0081	0.898	49.1	87.79%*	0.588
Gold	Commodity	0.0359	0.0189	0.891	47.3	66.76%**	0.836
Mean (all)	—	—	—	0.899	50.3	87.25%	0.779

[†] Post-calibration empirical coverage rate (target: 95%).

* EUR/USD: aleatoric uncertainty 89.7%; high irreducible noise limits calibration.

** Gold: aleatoric uncertainty 88.7%; episodic commodity shocks exceed bootstrap distributional assumptions.

Table 12 reports regime stability and economic value metrics.

Table 12. Cross-Asset Regime Stability and Economic Value (Test Set, 2022–2024).

Asset	RSI	Trans. Acc.	Epistemic (%)	PVR	VaR Hit Rate	Δ Sharpe
S&P 500	0.851	61.2%	31.1	+94.7%	1.32%	+0.017
NASDAQ-100	0.846	55.7%	30.9	+91.0%	0.53%	+0.011
Russell 2000	0.702	58.3%	17.2	+93.2%	2.65%	+0.017
EUR/USD	0.677	55.8%	10.3	+93.5%	1.78%	+0.110
Gold	0.713	56.0%	11.3	+93.8%	1.07%	+0.067

Three cross-cutting findings deserve emphasis. First, *predictive improvement is consistent across asset classes*: RADE-UQ reduces RMSE by 47–55% relative to HAR-RV across all five assets, with $R^2 > 0.88$ in every case. This uniformity confirms that heterogeneous ensembling captures volatility dynamics that transcend individual market microstructures.

Second, *cross-regime stability generalizes robustly*: the regime stability index ranges from 0.677 (EUR/USD) to 0.851 (S&P 500), and transition accuracy varies between 55.7% and 61.2%—consistently above the 30–40% benchmark of static models. The somewhat lower RSI for EUR/USD (0.677) and Russell 2000 (0.702) reflects the more abrupt regime transitions observed in currency and small-cap markets, rather than any architectural limitation of RADE-UQ.

Third, *economic value is consistently positive*: portfolio variance reduction exceeds 91% across all five assets, all VaR hit rates fall within the Basel III permissible range of 0.5–2.65%, and the Sharpe ratio improvement is positive for every asset. EUR/USD exhibits the largest Sharpe gain (+0.110), consistent with the established role of accurate volatility forecasting in FX carry-trade risk management.

The calibration challenge for EUR/USD (coverage 87.8%) and Gold (coverage 66.8%) reflects structurally high aleatoric uncertainty (89.7% and 88.7% respectively), driven by geopolitical shocks and supply disruptions that generate tail realizations outside the bootstrap empirical distribution. This finding is discussed as a structural limitation in Section 6.4.

6. Discussion

6.1. Interpretation of key findings

6.1.1. Heterogeneous ensembling drives performance gains

RADE-UQ's 55.6% RMSE increase is much higher than normal ensemble gains (15–30% range) reported in previous work (Christensen et al., 2023). The performance is driven by three complementary mechanisms:

Model complementarity: The combination of linear (HAR-RV), tree-based (random forest), and gradient-boosting (LightGBM) models captures a wide range of volatility dynamics. Parsimony allows HAR-RV to flourish in persistent low-volatility regimes. Random forest can manage nonlinear interactions and rapid leaps across deep trees. LightGBM uses gradient-based optimization to produce delicate prediction patterns. No single model dominates all situations; variety ensures resilience during regime transitions.

Feature richness and utilization: Our 42-dimensional feature space includes temporal, market, macroeconomic, and regime-specific data. Individual models may underuse subsets owing to architectural restrictions. Ensemble aggregation guarantees that all predicted signals are captured while addressing individual model shortcomings.

Bootstrap variance reduction: Training 50 iterations per model and averaging lowers overfitting when compared to individual occurrences. The bootstrap's RMSE increase of 2.9% (Table 5) validates this impact.

6.1.2. Exceptional adaptive capacity through multi-level mechanisms

An interval width quality of 0.962 indicates near-optimal adaptation (typical range: 0.3–0.6). Three methods create this capacity

Bootstrap diversity effects: During periods of high volatility, bootstrap samples show more variance because data disruptions enhance forecast disagreement, thereby expanding prediction ranges. During tranquil times, bootstrap samples converge, resulting in small intervals.

Ensemble heterogeneity: Model families react differently to volatile environments. Linear models generate small gaps during calm periods and broader intervals during stress. Tree-based models show opposing trends. Combining both results in a balanced adaptive response.

Calibration optimization: The post-hoc multiplier optimization ($c^* = 4.398$) guarantees that intervals scale adequately while keeping the 95% coverage aim. The huge multiplier indicates bootstrap variance underestimating of tail risk, which is a well-known problem that requires empirical adjustment.

6.1.3. Cross-regime stability via implicit adaptation

The regime stability index of 0.856 shows remarkable stability, despite our ablation result that explicit regime-conditional weights yield no advantage (Table 5). This resolves via three implicit adaptation channels:

Model-specific regime affinity: Different models inherently perform better in different regimes, even without explicit weighting. During periods of extreme volatility, LightGBM's gradient boosting responds quickly. HAR-RV's parsimony minimizes overfitting in low volatility scenarios. Equal weighting alters effective influence when relative accuracies vary.

Automatic feature importance adjustment: Tree-based models automatically modify feature relevance between samples. Jump indicators and leverage effects become more important during times of high volatility, while long-memory components take precedence during calm periods.

Bootstrap variance modulation: Bootstrap variance rises with uncertainty, naturally expanding intervals and minimizing overconfidence in extrapolation.

The 62.2% transition accuracy, which is nearly double the normal static model performance (30–40%), demonstrates actual adaptive capability during structural failures.

It is important to distinguish between two senses of "regime awareness" in RADE-UQ. The *explicit* regime channel—the Bi-LSTM RDN providing regime-conditional weights—currently adds no measurable predictive value, as confirmed by the ablation and the RDN's near-random classification accuracy (33.1%). The *implicit* regime channel—model-family affinity, automatic feature importance adjustment in tree-based models, and bootstrap variance modulation—produce the observed RSI of 0.856 and 62.2% transition accuracy without any regime labels. These two channels are architecturally independent: improving the explicit channel (the research directions of Section 6.5) would add on top of, not replace, the already-functioning implicit channel. The current framework thus delivers genuine regime-awareness at zero RMSE cost, while the explicit channel remains an open research opportunity.

6.1.4. Economic value and practical utility

Portfolio variance reduction of 94.6% requires cautious interpretation. The baseline (constant-volatility aiming) is intentionally poor. After friction corrections, the remaining variance reduction is 39.8%, calculated conservatively as 94.6% multiplied by $(1-0.30)$, $(1-0.20)$, and $(1-0.25)$. This remains economically meaningful; a 40% variance reduction amounts to around a 12% Sharpe ratio increase.

A VaR hit rate of 1.32%, which nearly matches the aim of 1%, supports the model's fitness for regulatory compliance. RADE-UQ's hit rate is within the Basel III mandated range of 0.5–1.5 percent.

The tiny sharpe enhancement (+0.015) and neutral utility gain (+0.00%) represent an essential reality: volatility forecasts primarily decrease risk, not boost profits. This corresponds with theory—volatility demonstrates predictability while returns are mostly unpredictable (Andersen et al., 2003).

6.2. The ensemble paradox: Why RADE-UQ despite LightGBM's superiority?

LightGBM alone obtains an RMSE of 0.0098, which is 29% better than the RADE-UQ ensemble (0.0138). This poses an important question: why accept the RMSE sacrifice? We offer three complementary justifications, supported by additional empirical evidence provided in the revision.

Justification 1: Uncertainty quantification imperative. LightGBM only delivers point predictions, not prediction intervals. Risk management applications demand reliable confidence bounds for VaR

calculations, regulatory compliance (Basel III), and portfolio optimization in the face of parameter uncertainty. RADE-UQ's bootstrap ensemble allows for full uncertainty quantification with statistical validation (Kupiec test, $p = 0.923$). The 41% RMSE sacrifice provides crucial functionality not found in single models, making it a reasonable trade-off for production deployment in risk-sensitive applications. This is analogous to a statistician accepting a wider confidence interval from a nonparametric test over a tighter but invalid interval from a model that violates distributional assumptions: interval validity takes precedence over interval width.

Justification 2: Robustness and temporal stability. Beyond average RMSE, production deployment requires stable, predictable performance across market conditions. Table 6 demonstrates that LightGBM alone exhibits substantial temporal instability: RMSE ranges from 0.0089 to 0.0161 across quarterly sub-periods of the test set (CV = 18.4%). This large variability reflects the well-known tendency of gradient boosting to overfit regime-specific patterns in the training data, leading to sharp performance degradation when market conditions shift. RADE-UQ's ensemble diversification substantially mitigates this risk: RMSE ranges from 0.0132 to 0.0151 (CV = 6.3%), a 2.9 $\times$ reduction in temporal variability. For a risk management system that must perform reliably through market cycles, the stability premium justifies the aggregate RMSE cost. If core assumptions are broken, single models may collapse catastrophically. Ensembles mitigate model risk: if LightGBM suffers with a regime transition, HAR-RV and Random Forest retain baseline performance.

Justification 3: Interpretability and regulatory acceptance. Production risk management systems in financial institutions require regulatory approval. Models that include well-understood econometric components (HAR-RV) alongside gradient boosting are significantly easier to validate and explain to risk committees and regulators than pure black-box models. The HAR-RV inclusion provides an interpretable anchor. Production deployment needs support from risk committees and regulators, who frequently regard black-box gradient boosting with skepticism.

Table 13. Practical model selection guide: LightGBM vs. RADE-UQ.

Requirement	LightGBM	RADE-UQ
Lowest aggregate RMSE	✓	
Point forecast only	✓	
Academic benchmarking	✓	
Statistically valid UQ		✓
Basel III VaR compliance		✓
Calibrated 95% prediction intervals		✓
Temporal stability (low RMSE CV)		✓
Regulatory interpretability		✓
Regime-robust deployment		✓
Cross-asset generalization		✓

✓: primary strength of this model for the stated requirement.

Practical decision rule. Table 13 provides a structured guide for model selection based on application requirements. LightGBM is the appropriate choice when pure point accuracy is the primary objective—for example, academic benchmarking, research competitions, or applications with no regulatory oversight. RADE-UQ is the appropriate choice when the application requires any of the

following: statistically valid prediction intervals, regulatory compliance (Basel III VaR backtesting), production stability across market regimes, or stakeholder interpretability through the inclusion of a recognised econometric model.

The RMSE sacrifice (0.0098, $\rightarrow$ 0.0138, +41%) is the price of the capabilities listed in the lower panel of Table 13. For any application requiring even one of those capabilities, the sacrifice is justified. For applications requiring only raw point accuracy, LightGBM is preferable, and we explicitly recommend it.

6.3. Uncertainty decomposition and improvement pathways

The uncertainty composition ratio (UCR) of 0.362 indicates that 36.2% of overall uncertainty is epistemic (reducible through model improvements), and 63.8% is aleatoric (irreducible data noise). This breakdown indicates both potential and limitations

Epistemic uncertainty reduction pathways: Three techniques to decrease model-related uncertainty: (1) Extending temporal coverage beyond 10 years results in more diversified regime experiences (estimated 10–15% epistemic loss). (2) Enhanced features—alternative macroeconomic indicators or market microstructure variables could capture additional signals (estimated 5–10% reduction). (3) Architectural sophistication—deeper networks or Transformer architectures may extract complex patterns, but complexity increases overfitting risk (estimated 5–10% reduction with high uncertainty).

Optimally, these enhancements might reduce epistemic uncertainty by 20–35%, resulting in a UCR of 0.23–0.29. However, further model improvements may result in decreasing returns.

Aleatoric uncertainty as a fundamental limit: The 63.8% aleatoric component includes (1) Proxy error: realized volatility is a noisy estimate of genuine latent volatility (10–20% contribution); (2) Genuine stochasticity refers to unanticipated news shocks and irrational behavior (40–50%), which is the greatest forecasting limit.

Practical implication: Users should calibrate their expectations. Even with flawless models, 60–70% uncertainty remains irreducible, implying forecast accuracy has a realistic upper bound.

6.4. Limitations

Despite strong performance, our study has five important limitations

Limitation 1: Regime misclassification and incomplete RDN integration. The Bi-LSTM regime detection network achieves a validation accuracy of 34.7%, which is close to the random baseline of 33.3% for a three-class classification problem. This limitation—already disclosed in the original submission—means that the regime detection component currently adds little discriminative power; the ablation study confirms zero RMSE improvement from regime-conditional weights (Table 4). The root cause is the supervised training scheme: K-means labels derived from realized volatility introduce regime boundary noise, particularly during transition periods where volatility levels are ambiguous. We propose two concrete remediation strategies: (i) *Self-supervised pre-training* of the Bi-LSTM on the volatility time series as a sequence prediction task before fine-tuning on regime labels, which has demonstrated 15–25% accuracy improvements in analogous NLP settings; and (ii) replacement of discrete three-class labels with *continuous regime probability vectors* derived from hidden markov models, enabling smooth regime transitions rather than hard boundaries. The current validation set ($n = 377$) is likely insufficient for reliable three-regime weight optimization; a minimum of 500–600

observations per regime is recommended for future work.

Root-cause analysis. The current supervised training scheme uses K-means clustering on realized volatility to assign regime labels. K-means partitions the volatility space into three clusters by minimizing within-cluster variance, but the resulting boundaries do not necessarily correspond to stable or economically meaningful regimes. In particular: (i) K-means regime boundaries are sensitive to the choice of K and initialization; (ii) cluster assignment is hard (each observation belongs to exactly one regime), whereas true market regimes transition gradually; and (iii) realized volatility at the cluster boundary is ambiguous—the same level of volatility may represent different regimes depending on whether the market is entering or exiting a stress period. These sources of label noise reduce the Bi-LSTM’s ability to learn generalizable regime representations, leading to test accuracy near the random baseline.

Three remediation strategies are proposed and incorporated into the research roadmap of Section 6.5

- *Self-supervised pre-training:* train the Bi-LSTM on the volatility sequence prediction task before fine-tuning on regime labels. This approach, standard in NLP (Devlin et al., 2019), has demonstrated 15–25% classification accuracy improvements in analogous time series settings, which would bring RDN accuracy to the 50–60% range where regime-conditional weighting is expected to yield measurable RMSE gains.
- *Continuous regime indicators via HMM:* replace the discrete three-class K-means labels with continuous regime probability vectors from a hidden markov model, allowing smooth regime transitions and eliminating hard classification boundaries. This removes the label noise problem at its source.
- *Larger validation set:* the current regime-conditional weight optimization uses $n = 377$ validation observations divided across three regimes, yielding approximately 126 observations per regime. Reliable regime-specific weight estimation requires a minimum of 500–600 observations per regime; an expanded validation window covering two full market cycles is recommended.

Limitation 2: Calibration precision trade-off. Achieving 95% coverage requires a multiplier $c^* = 4.398$ (vs. the theoretical 1.96), which widens intervals by 124%. Alternative calibration approaches, such as conformal prediction (Vovk et al., 2005) and quantile regression, may achieve greater balance.

Failure mode category 2: Feature over-reliance during structural breaks. The 42-dimensional feature space confers predictive richness but also creates a risk that tree-based models learn spurious correlations during the training period (2014–2021) that do not persist after structural shifts, such as the post-pandemic inflation regime or the rapid Federal Reserve tightening cycle. RADE-UQ mitigates this risk through two mechanisms: (1) the bootstrap ensemble’s variance-averaging across $B = 50$ iterations reduces the influence of any individual spurious signal; and (2) the minimum weight constraint ($w_{\min} = 0.05$) prevents the triple-adaptive weighting from concentrating entirely on a single model family that may have overfitted to regime-specific patterns. The subperiod analysis (Table 13) confirms stable performance across quarterly windows (RMSE CV = 6.3%), suggesting that feature over-reliance is currently managed adequately. Nevertheless, practitioners deploying in rapidly shifting macro environments should consider periodic feature importance audits using SHAP values to identify and down-weight any features whose predictive content has degraded.

Failure mode category 3: Calibration Multiplier Instability Under Distribution Shift. The large post-hoc calibration multiplier ($c^* = 4.398$, approximately 2.25 times the theoretical Gaussian value of

1.96) indicates that the bootstrap systematically underestimates tail risk in the validation set. This is a known limitation of bootstrap-based prediction intervals, which undercover in the tails when the data-generating distribution exhibits fat tails or regime-switching behavior (Breiman, 1996). The multiplier is optimized on the validation set (2021-2022); if the volatility distribution shifts structurally—as occurred during the 2022 rate-hiking cycle—the stored multiplier may no longer achieve 95% coverage on new data. We recommend re-calibrating c^* on a rolling 12-month validation window at each re-estimation, adding approximately one additional binary search call (13 iterations, negligible computational cost) to the quarterly updating procedure described in Appendix C.5.

Limitation 3: Calibration under high aleatoric regimes. Cross-asset validation confirms RADE-UQ's generalizability across equity indices, foreign exchange, and commodities. However, bootstrap-based calibration faces structural limitations for assets with high aleatoric uncertainty. EUR/USD (aleatoric: 89.7%) and Gold (aleatoric: 88.7%) achieve post-calibration coverage of 87.8% and 66.8%, respectively, despite optimized multipliers ($c^* = 4.685$ and 4.966). The root cause is that episodic shocks in currency and commodity markets—geopolitical events, supply disruptions, central bank interventions—generate tail realizations that exceed the empirical distribution of bootstrap samples drawn from the training set. Alternative calibration approaches, including conformal prediction (Vovk et al., 2005) and quantile regression forests, may achieve better coverage in such regimes and are identified as a priority for future research.

Limitation 4: Computational cost. Training 153 models in 48.2 hours is extensive. Production deployment need efficient updates. Cost-cutting measures might include incremental learning, GPU acceleration, and model compression.

Limitation 5: Economic value assumptions. The estimates for portfolio variance reduction assume frictionless trade. Real-world implementation involves transaction costs, slippage, and market effect. Our number of 94.6% is an upper bound; realistic estimates are 30-50% after frictions.

Limitation 6: Transferability to less liquid markets. The five asset classes evaluated in this study all trade on deep, liquid markets: U.S. equity indices benefit from tight bid-ask spreads, continuous price discovery, and minimal transaction costs; EUR/USD is the most liquid currency pair globally; Gold futures trade on regulated exchanges with high turnover. Deployment of RADE-UQ in less liquid contexts—such as emerging market equity indices, illiquid commodity contracts, or corporate bond volatility—would require three adaptations.

First, *feature set adjustment*: the macroeconomic conditioning variables from FRED (VIX, unemployment, federal funds rate, term spread, CPI, credit spread) are U.S.-centric and may have limited predictive power for emerging market volatility. Local equivalents—such as country-specific policy rates, sovereign credit default swap (CDS) spreads, and commodity-specific supply indicators—should replace or supplement the current FRED variables. The nine regime-specific features (Appendix A.1) are market-agnostic and transfer without modification.

Second, *recalibration of the multiplier c^** : the calibration multiplier must be re-estimated on an asset-specific validation set, as the tail behavior of realized volatility differs substantially across markets. The binary search algorithm (Algorithm 1, Appendix B.2) is computationally inexpensive and can be re-run for any new asset within minutes.

Third, *economic value reassessment*: the portfolio variance reduction figures (91-95%) assume transaction cost structures comparable to the assets evaluated. In less liquid markets, higher bid-ask spreads and market impact costs would substantially reduce realizable gains. We recommend applying

the friction adjustment framework of Fleming et al. (2001, 2003)—already used in our economic value analysis—with asset-specific transaction cost estimates.

6.5. Future research directions

We identify seven promising research avenues

(1) **Advanced regime-adaptive mechanisms with improved RDN training.** The Bi-LSTM RDN's current test accuracy of 33.1%-near the random baseline of 33.3%-represents the primary bottleneck for the explicit regime-conditioning component of RADE-UQ. Three concrete improvement pathways are proposed, in order of implementation priority:

Priority 1 - Self-supervised pre-training (estimated 12-16 weeks implementation): Pre-train the Bi-LSTM as a volatility sequence predictor before fine-tuning on regime labels. This leverages the unlabeled time series structure and is expected to improve accuracy to 50–60%, the threshold at which regime-conditional weights outperform equal weighting in simulation studies.

Priority 2 - Continuous HMM regime indicators (estimated 8–12 weeks): replace K-means hard labels with Hidden Markov Model regime probability vectors $p_t \in [0, 1]^K$. This eliminates the label noise problem at its source and enables smooth, probabilistic weighting rather than hard regime assignment.

Priority 3 - Expanded validation set (data collection): collect at least 1500–1800 validation observations (approximately 6–7 years) to ensure 500–600 observations per regime for reliable weight optimization. This requires extending the dataset to pre-2014 data, for which high-quality FRED macroeconomic series are available.

(2) **Online learning.** Apply online gradient descent to ensemble weights, updating them as fresh data enters.

(3) **Multivariate extension.** Extend to covariance matrix forecasting, where obstacles include dimensionality and positive definiteness restrictions.

(4) **High-frequency realized volatility as target variable.** Replacing the daily-frequency rolling standard deviation proxy with intraday high-frequency realized variance (5-minute RV) constitutes the most impactful near-term extension of RADE-UQ. Three specific improvements are expected: (i) *Reduced proxy noise*: intraday RV provides a lower-noise, asymptotically unbiased estimate of integrated variance, which would reduce the calibration multiplier c^* (currently 4.4 due to partial bootstrap underestimation of tail risk from noisy targets) and improve interval coverage for EUR/USD and Gold; (ii) *Richer features*: the 42-feature set could be augmented with realised bipower variation, jump components estimated from intraday data, and microstructure noise measures, all of which have established predictive content in the high-frequency RV literature (Barndorff-Nielsen and Shephard, 2002; Andersen et al., 2001); (iii) *Intraday forecasting*: extending the forecast horizon to 1-hour or end-of-day intraday targets using tick-level data, relevant for high-frequency trading risk management.

(5) **Alternative UQ methods.** Compare bootstrap to bayesian neural networks, Monte Carlo dropout, and conformal prediction.

(6) **Interpretability enhancement.** To explain predictions, include SHAP values or an attention visualization.

(7) **Live deployment study.** Collaborate with a financial institution to assess realized economic value.

The 36.2% epistemic uncertainty indicates significant improvement potential. The 63.8% aleatoric uncertainty establishes realistic expectations, since even flawless models cannot remove market unpredictability.

(8) *Extension to emerging markets and less liquid asset classes.* Adapting RADE-UQ to emerging market equity indices (e.g., MSCI EM, BRIC indices), illiquid commodity contracts (e.g., agricultural futures, rare earth metals), and fixed income volatility (e.g., MOVE index, sovereign bond markets) constitutes a natural next step. The primary research questions are: (i) which of the 42 features retain predictive content in lower-liquidity, higher-friction environments; (ii) whether the bootstrap calibration procedure achieves adequate coverage under the fatter-tailed volatility distributions characteristic of emerging markets; and (iii) how to adapt the economic value framework when transaction cost structures differ substantially from liquid exchange-traded markets.

7. Conclusions

This study offers RADE-UQ (regime-adaptive deep ensemble with uncertainty quantification), a complete system that addresses three important gaps in volatility forecasting: the lack of principled uncertainty quantification, static models in dynamic markets, and insufficient assessment frameworks. RADE-UQ achieves cutting-edge performance by combining regime detection via Bi-LSTM, bootstrap-based uncertainty quantification across heterogeneous models, and triple-adaptive ensemble weighting with statistically validated prediction intervals, which are critical for risk-sensitive applications.

7.1. Summary of contributions

Our empirical evaluation using 10 years of S&P 500 data demonstrates four principal contributions

C1: State-of-the-art predictive performance across asset classes. RADE-UQ achieves an RMSE of 0.0138 on S&P 500, demonstrating a 55.6% improvement over the HAR-RV baseline, with R^2 improvement from 0.602 to 0.922. Cross-asset validation on four additional markets (NASDAQ-100: +47.7%; Russell 2000: +52.0%; EUR/USD: +49.1%; Gold: +47.3%) confirms consistent improvements of 47–55% over HAR-RV $R^2 > 0.89$ across all assets, establishing RADE-UQ as a generalizable framework rather than a single-asset benchmark. This outperforms current machine learning benchmarks (Christensen et al., 2023) while preserving computational efficiency (48.2 hours of training, 0.42 seconds of inference).

C2: Rigorous uncertainty quantification. Kupiec's test validates the post-calibration empirical coverage, which is 94.97% (target: 95%). An interval width quality of 0.962 indicates outstanding adaptation potential, as intervals extend 8–12 times during stress periods. To our knowledge, this is the first volatility forecasting research to achieve high point accuracy, well-calibrated intervals, statistical validation, and adaptive interval widths.

C3: Comprehensive evaluation framework. We present 18 measures for calibration quality, uncertainty decomposition (epistemic: 36.2%, aleatoric: 63.8%), regime-specific performance (RSI: 0.856, transition accuracy: 62.2%), and economic value. The revised manuscript extends this framework with crisis-period evaluation (Table 7), reporting that calibrated coverage remains above 90% during the COVID-19 shock, 2022 rate-hiking cycle, and 2023 banking stress, and that interval width quality increases to 1.38–1.41 during these episodes-confirming adaptive behaviour precisely when it is most needed. This holistic evaluation gives new insight into model behavior.

C4: Demonstrated economic value. The portfolio variance reduction of 94.6%, VaR hit rate of 1.32% (target: 1%), and Sharpe improvement of +0.015 demonstrate practical value. Even with modest friction assumptions (50% degradation), a 40% decrease in residual variance leads to significant Sharpe ratio benefits.

7.2. Methodological insights

Beyond specific performance gains, this study yields three methodological insights

Insight 1: Heterogeneous ensembles excel. Combining model families improves resilience and uncertainty calibration over homogenous ensembles. Model complementarity captures the many volatility dynamics; no single design dominates all market situations.

Insight 2: Uncertainty quantification is essential. Most forecasting studies only provide point forecasts, but risk management applications require prediction intervals. Bootstrap ensembles give principled and confirmed uncertainty estimates at a low computational cost.

Insight 3: Comprehensive evaluation reveals trade-offs. Traditional RMSE-focused assessments overlook key features. LightGBM alone yields the highest point accuracy, but an ensemble offers improved robustness and uncertainty quantification. The optimal model choice is determined by the application's objectives.

The temporal stability analysis (Table 6) reinforces this insight quantitatively: LightGBM's coefficient of variation of RMSE across quarterly sub-periods (18.4%) is nearly three times larger than RADE-UQ's (6.3%). A practitioner relying solely on aggregate RMSE would select LightGBM and unknowingly accept a model that underperforms substantially in certain market regimes. The ensemble's stability premium is invisible to standard evaluation protocols but critical for production deployment.

7.3. Practical implications

For financial institutions, RADE-UQ offers actionable benefits

Risk management. Calibrated 95% intervals for accurate VaR and CVaR calculations in accordance with Basel III. Adaptive interval widths (IWQ: 0.962) provide accurate risk estimation across market situations.

Portfolio optimization. Improved volatility forecasting improves mean-variance allocation. Even with cautious friction modifications, a variance decrease of 47% indicates significant Sharpe ratio increases.

Derivatives pricing. More accurate volatility projections lower option pricing mistakes and hedging costs, especially for longer-dated options.

Regime-aware trading. The high RSI (0.856) and transition accuracy (62.2%) allow for proactive risk management during structural breakdowns.

Caveat on regime detection. The cross-regime stability reported above (RSI: 0.856, transition accuracy: 62.2%) is achieved through implicit adaptation mechanisms rather than the explicit Bi-LSTM RDN, whose classification accuracy (33.1%) is currently near the random baseline. Practitioners should not rely on the RDN's regime probability outputs for trading signal generation in the current implementation. The regime-awareness benefit is real but derives from the heterogeneous ensemble's model-family diversity rather than from learned regime labels. Improved RDN performance—following the pathways described in Section 6.5—would unlock additional regime-conditioning benefits currently unavailable.

7.4. Limitations and future directions

Users should be mindful of five key limitations: (1) Transferability to less liquid markets: the five assets evaluated (S&P 500, NASDAQ-100, Russell 2000, EUR/USD, Gold) all trade on deep, liquid markets. Deployment in emerging market equities or illiquid commodity contracts would require

feature set adaptation, asset-specific recalibration of c^* , and reassessment of economic value under higher transaction costs; **(2)**: 48-hour training requires efficient update procedures; **(3)**: Economic value upper bounds—94.6% assumes frictionless trading (realistic: 30–50%); **(4)**: Incomplete regime integration—explicit regime-conditional weights showed no gain; **(5)**: Calibration cost—achieving 95% coverage requires 124%.

Promising research directions include advanced regime-adaptive mechanisms with continuous indicators, online learning for real-time adaptation, multivariate extension to covariance forecasting, high-frequency intraday prediction, alternative uncertainty quantification methods, interpretability enhancement via SHAP values, and live deployment studies measuring realized economic value.

The 36.2% epistemic uncertainty indicates significant improvement potential. The 63.8% aleatoric uncertainty establishes realistic expectations, since even flawless models cannot remove market unpredictability.

7.5. *Final remarks*

RADE-UQ demonstrates that well-designed ensemble frameworks that combine traditional econometric models with modern machine learning, supplemented by rigorous uncertainty quantification, can achieve cutting-edge performance while retaining interpretability, computational efficiency, and practical utility. The framework's high adaptive capability (IWQ: 0.962), cross-regime stability (RSI: 0.856), and economic value (PVR: 94.6%, conservatively 40%) make it a reliable option for forecasting output volatility.

The horizon of volatility forecasting goes beyond simple point accuracy. Comprehensive uncertainty quantification, statistical validation, adaptive capability, and economic value evaluation are all equally important—dimensions that are sometimes overlooked in machine learning literature yet required for real-world deployment. Future study should use this multidimensional approach, evaluating forecasting systems holistically using end-user objectives rather than limited accuracy criteria alone. Only by doing such a thorough examination will we be able to develop forecasting systems that really meet the demands of practitioners while also enhancing scientific rigor in financial economics.

Author contributions

Khalid Jeaab: Conceptualization, Methodology, Software, Formal analysis, Data curation, Visualization, Writing – original draft, Writing – review and editing.

Smaaine Ouaharahe: Validation, Writing – review and editing, Supervision.

Youness Saoudi: Validation, Writing – review and editing, Analysis, Supervision.

All authors have read and approved the final version of the manuscript for publication.

Use of AI tools declaration

The authors declare they have not used Artificial Intelligence (AI) tools in the creation of this article.

Conflict of interest

All authors declare no conflicts of interest in this paper.

References

- Andersen TG, Bollerslev T, Diebold FX, et al. (2001) The distribution of realized stock return volatility. *J Financ Econ* 61: 43–76. [https://doi.org/10.1016/S0304-405X\(01\)00055-1](https://doi.org/10.1016/S0304-405X(01)00055-1)
- Andersen TG, Bollerslev T, Diebold FX, et al. (2003) Modeling and forecasting realized volatility. *Econometrica* 71: 579–625. <https://doi.org/10.1111/1468-0262.00418>
- Bahdanau D, Cho K, Bengio Y (2015) Neural machine translation by jointly learning to align and translate. In: *International Conference on Learning Representations*.
- Barndorff-Nielsen OE, Shephard N (2002) Econometric analysis of realized volatility and its use in estimating stochastic volatility models. *J R Stat Soc B* 64: 253–280. <https://doi.org/10.1111/1467-9868.00336>
- Bollerslev T (1986) Generalized autoregressive conditional heteroskedasticity. *J Econometrics* 31: 307–327. [https://doi.org/10.1016/0304-4076\(86\)90063-1](https://doi.org/10.1016/0304-4076(86)90063-1)
- Breiman L (1996) Bagging predictors. *Mach Learn* 24: 123–140. <https://doi.org/10.1023/A:1018054314350>
- Breiman L (2001) Random forests. *Mach Learn* 45: 5–32. <https://doi.org/10.1023/A:1010933404324>
- Cho K, Van Merriënboer B, Gulcehre C, et al. (2014) Learning phrase representations using RNN encoder-decoder for statistical machine translation. In: *Conference on Empirical Methods in Natural Language Processing*, 1724–1734.
- Christensen K, Siggaard M, Veliyev B (2023) A machine learning approach to volatility forecasting. *J Financ Econometrics* 21: 1680–1727. <https://doi.org/10.1093/jjfinec/nbac020>
- Christoffersen PF (1998) Evaluating interval forecasts. *Int Econ Rev* 39: 841–862. <https://doi.org/10.2307/2527341>
- Corsi F (2009) A simple approximate long-memory model of realized volatility. *Jo Financ Econometrics* 7: 174–196. <https://doi.org/10.1093/jjfinec/nbp001>
- Fleming J, Kirby C, Ostdiek B (2001) The economic value of volatility timing. *J Financ* 56: 329–352.
- Fleming J, Kirby C, Ostdiek B (2003) The economic value of volatility timing using “realized” volatility. *J Financ Econ* 67: 473–509. [https://doi.org/10.1016/S0304-405X\(02\)00259-3](https://doi.org/10.1016/S0304-405X(02)00259-3)
- Gneiting T, Raftery AE (2007) Strictly proper scoring rules, prediction, and estimation. *J Am Stat Assoc* 102: 359–378. <https://doi.org/10.1198/016214506000001437>
- Hamilton JD (1989) A new approach to the economic analysis of nonstationary time series and the business cycle. *Econometrica* 57: 357–384. <https://doi.org/10.2307/1912559>
- Hansen PR, Lunde A (2005) A forecast comparison of volatility models: Does anything beat a GARCH(1,1)? *J Appl Econometrics* 20: 873–889. <https://doi.org/10.1002/jae.800>
- Hochreiter S, Schmidhuber J (1997) Long short-term memory. *Neural Comput* 9: 1735–1780. <https://doi.org/10.1162/neco.1997.9.8.1735>
- Ke G, Meng Q, Finley T, et al (2017) LightGBM: A highly efficient gradient boosting decision tree. *Adv Neur Inf Process Syst* 30: 3146–3154.
- Kupiec PH (1995) Techniques for verifying the accuracy of risk measurement models. *J Deriv* 3: 73–84.

- Patton AJ (2011) Volatility forecast comparison using imperfect volatility proxies. *J Econometrics* 160: 246–256. <https://doi.org/10.1016/j.jeconom.2010.03.034>
- Vaswani A, Shazeer N, Parmar N, et al. (2017) Attention is all you need. *Adv Neur Inf Process Syst* 30: 5998–6008.
- Vovk V, Gammerman A, Shafer G (2005) *Algorithmic Learning in a Random World*. Springer.
- West KD (1996) Asymptotic inference about predictive ability. *Econometrica* 64: 1067–1084. <https://doi.org/10.2307/2171956>
- Zhou ZH (2012) *Ensemble methods: Foundations and algorithms*.



AIMS Press

©2026 the Author(s), licensee AIMS Press. This is an open access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0>)