

Research Article

Statistical and Economic Comparisons of Econometric, Linear, and Machine Learning Models

Suganya Eliot^{1,*}, Muthulakshmi Angamuthu²¹ Department of Information Technology and Systems, Victorian Institute of Technology, Melbourne, Victoria, Australia.² Department of Mathematics, PSG College of Technology, Coimbatore, Tamil Nadu, India.

ARTICLE INFO

Article History

Received 08 Jul 2026

Revised 20 Aug 2026

Accepted 15 Sep 2026

Published 06 Oct 2026

Keywords

ETF volatility,
EGARCH,
GARCH,
machine learning,
XGBoost,
LightGBM,
forecast comparison,
QLIKE,
volatility targeting.

ABSTRACT

This study compares econometric, linear regression, and machine-learning models for forecasting five-day close-to-close volatility across 12 diversified US exchange-traded funds. The dataset covers 2,511 trading days from September 2016 to August 2026 and includes equities, bonds, precious metals, commodities, and real estate. Fourteen models were evaluated using a leakage-free expanding-window framework, with quarterly retraining and hyperparameter tuning based only on training data. Performance was assessed using forecast errors, QLIKE, out-of-sample R-squared, statistical tests, the Model Confidence Set, regime analysis, SHAP explanations, and volatility-targeting outcomes. EGARCH achieved the best average QLIKE (0.382) and model rank (1.92), improving QLIKE by 12.2% over historical volatility and 5.3% over GARCH. The ensemble produced the lowest MAE (0.0412) and RMSE (0.0612) and the highest out-of-sample R-squared (0.159), while HAR achieved the highest volatility-targeting Sharpe ratio (0.895). Model rankings differed significantly ($p < 0.001$). Although EGARCH generally performed strongly, it did not significantly outperform GARCH or GJR-GARCH. Overall, machine-learning models did not consistently outperform econometric approaches, and the best model depended on the asset class, market conditions, evaluation measure, and forecasting purpose.

1. INTRODUCTION

Volatility forecasts determine the amount of leverage, limit of risks, hedges, portfolios construction, and capital allocation. Exchange-traded funds (ETFs) provide a suitable test bed for the comparison of forecasting methods because they give access to several asset classes that demonstrate different behaviour in return distribution. The response of Equity volatility to positive and negative market shocks is different. Whereas Treasury bonds, corporate bonds, commodities, gold, and real estate are influenced by factors such as interest rates, inflation, and market liquidity. Hence a model that suits a stock index may not suit for assets in a diversified portfolio.

Econometric models model the time-varying variance of the returns. ARCH and GARCH models incorporate volatility persistence over time [1], [2], GJR-GARCH model allows negative shocks to affect volatility differently from positive [3], EGARCH model works in logarithms but also captures an asymmetric effect of shocks [4]. Exponential weighting is popular because of the simplicity of implementation, while heterogeneous autoregressive models capture the effect of different lags of volatility without additional parameters [5]. Regularized regression, random forests, and gradient-boosted trees provide the alternative flexible methods to link predictors to future volatility [6], [7], [8] — however, their superiority does not mean better out-of-sample performance. The comparison of average errors of the different models alone does not suffice the purpose. Volatility measure and evaluation metrics are also factors that influence the model rankings. Hence forecasting accuracy, statistical significance and economic usefulness are considered in this study. It also examines whether model rankings change across assets and volatility regimes. SHAP values are used to explain the predictions of the tree-based models.

Hereinafter, leakage-controlled cross-assets evaluation of econometric, regularised-linear, tree-based and ensemble approaches to overlapping multi-day volatility forecasts for ETFs will be conducted. The reproducible framework is implemented for all 12 ETFs. The results of QLIKE evaluation are treated separately from regular level-based measures. Statistical uncertainty is considered with the help of Hansen-Lunde-Nason Model Confidence Set, and model performance is linked with the volatility targeting strategy accounting for transaction costs. Additional analyses examine later evaluation periods, non-overlapping forecast horizons, different ETF categories, and alternative assumptions about transaction costs

*Corresponding author. Email: eliotsuganya@gmail.com

and leverage. There is no universal winner: asymmetric econometric models turn out to perform better according to QLIKE, while forecast combination and persistence-based models are preferable when assessed through other criteria and investment strategy.

2. LITERATURE REVIEW

2.1 Conditional Volatility Models

The idea that return variance depends on time through the past shocks is formulated in Engle's [1] ARCH model. Bollerslev [2] further enhanced the model by including lagged conditional variance into the equation, thereby developing GARCH model. It is a benchmark for persistence in volatility. Asymmetric reactions of financial returns to the shock of the same magnitude (bad news increases volatility more than good news of the same size) is a property of financial returns. This phenomenon is accounted for in the case of GJR-GARCH model with the help of indicator for negative innovations [3], and in the case of EGARCH model, where the logarithm of variance is used and the impact of positive and negative shocks may differ [4]. Hansen and Lunde [9] demonstrated the high persistence of the GARCH (1,1) model making it a benchmark against which any new models must be compared.

2.2 Linear and machine learning forecasts

Linear models remain valuable as their coefficients are easier to interpret. Corsi's [5] HAR model describes the persistence in terms of daily, weekly and monthly components. Ridge, Lasso and Elastic Net address the problem of multicollinearity and large number of predictors using different kinds of regularisation [10], [11], [12]. Tree-based models, in turn, can catch nonlinearities, interactions and threshold effects without specifying them in advance. Random forests reduce the forecast variance by averaging the predictions from several trees trained on the bootstrap samples of data and predictors [6]. XGBoost and LightGBM use the gradient boosting technique with added regularisation and optimisation [7], [8]. The effectiveness of this approach, however, still must be proved in chronological out-of-sample forecasting.

2.3 Evaluation and research gap

As volatility cannot be observed, model ranking can depend strongly on the choice of proxy and loss function. Patton [13] derives general principles of evaluation of the forecasts when the variable itself is evaluated with the error. QLIKE measure is suitable for this task as it punishes severely for the underpredictions and is appropriate for variance forecast. The Diebold–Mariano test examines whether two models have equal forecasting accuracy [14]. The Model Confidence Set identifies models whose performance cannot be statistically distinguished from the best-performing model [15]. Validation techniques must respect the time-based sequence of dependent data [16]. Expanding on the previous research, this study integrates statistical testing, market-regime analysis, model interpretation, and an assessment of economic utility.

2.4 Research questions

RQ1: Which model class delivers the most accurate out-of-sample five-day ETF volatility forecasts?

RQ2: Is there any statistical significance of the forecasting accuracy differences between ETFs?

RQ3: Is performance different in various asset classes and volatility levels?

RQ4: What market and technical features affect machine learning forecasts?

RQ5: Are statistical gains really translated to practical results of volatility targeting?

3. DATA AND METHODS

3.1 Data and Assets

The daily adjusted closing price and volume data have been retrieved from the provided Yahoo Finance export file. The two-line header from the yfinance has been removed, observations have been sorted by date, duplicate dates have been identified and the common calendar with the full set of prices for all ETFs has been kept. The incomplete observation of 28 August 2026 has been deleted. High, low and opening prices have not been used. The final panel includes 2,511 days in which all ETFs trade, from 1 September 2016 until 31 August 2026. Since the target depends on the next five trading days returns, the last five observations cannot be included in either estimation or evaluation of models.

TABLE I. SAMPLE COMPOSITION AND FULL-PERIOD RETURN CHARACTERISTICS

ETF	Exposure	Return	Volatility	Skewness	Excess kurtosis
AGG	Aggregate bonds	0.014	0.054	-1.13	19.19
DBC	Commodities	0.096	0.180	-0.76	4.49
EEM	Emerging-market equity	0.082	0.209	-0.73	9.43
EFA	Developed-market equity	0.090	0.171	-0.96	15.68
GLD	Gold	0.119	0.164	-0.73	8.20
IWM	US small-cap equity	0.100	0.231	-0.71	8.92
LQD	Corporate bonds	0.020	0.087	0.29	27.79
QQQ	US growth equity	0.189	0.226	-0.38	7.57
SHY	Short Treasury bonds	0.017	0.016	0.68	8.71
SPY	US large-cap equity	0.143	0.180	-0.60	15.26
TLT	Long Treasury bonds	-0.025	0.148	0.02	5.20
VNQ	Real estate	0.048	0.210	-1.56	26.25

Annualised returns and volatility are based on daily log returns. Kurtosis is expressed as excess kurtosis. Negative skewness in VNQ, along with high excess kurtosis in LQD, VNQ, AGG, EFA, and SPY suggests that returns have a distribution far from normal. These characteristics imply that heavy-tailed and asymmetric models for conditional volatility should be considered in the study.

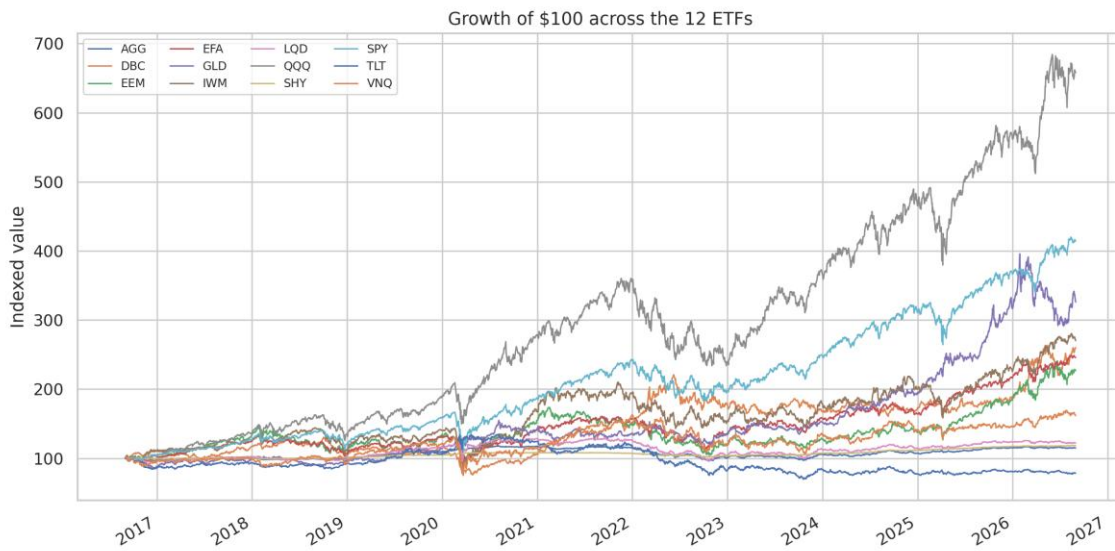


Fig.1. Growth of 100 monetary units across the ETF sample

3.2 Target and Predictors

For ETF i , the daily logarithmic return is calculated as:

$$r_{i,t} = \ln\left(\frac{P_{i,t}}{P_{i,t-1}}\right),$$

where $P_{i,t}$ represents the adjusted closing price on day t . The main forecasting target is the annualised close-to-close volatility over the following five trading days:

$$RV_{i,t+1:t+5} = \sqrt{\frac{252}{5} \sum_{j=1}^5 r_{i,t+j}^2}$$

A five-day volatility measure is more stable as compared to a single squared daily return as weekly review happens with most portfolio and risk-management decisions in practice. This is a close-to-close volatility measure as compared against the variance obtained from intraday prices.

Each predictor is built using only available data until the end of day t . The feature set is composed of current and lagged returns, cumulative returns over five and 22 days, negative return indicator, total and downside volatilities over 5, 10, 22, and 44 days, volatility-of-volatility, largest absolute return over the past 22 days, price-to-moving-average ratios, RSI, scaled MACD, and changes and standardised values of log trading volume. Returns and volatility statistics for SPY, TLT, GLD, DBC, and VNQ represent broader market environment. Cross-sectional volatility is calculated for all other ETFs except the ETF under consideration. There is no transformation that makes use of forward-looking data.

3.3 Models

A total of 14 forecasts is compared. Historical 22-day volatility and an exponential weight moving average with $\lambda = 0.94$ are used as simple benchmarks. The HAR-style regression forecasts are constructed based on logarithm of 5-, 22-, and 44-day close-to-close volatilities. As the dataset consists of daily closing prices, the approach suggested by Corsi [5] to predict intraday realised volatility is adapted to close-to-close scenario. Econometric models comprise Student-t GARCH (1,1), GJR-GARCH(1,1), and EGARCH(1,1). Their parameters are re-estimating each quarter, while in-between returns are updated and parameters are fixed. The average five-step forecasts of conditional variances are aggregated and annualised.

Linear forecasts are obtained by OLS, Ridge, Lasso, and Elastic Net methods. Machine learning forecasts are produced by Random Forest, XGBoost, and LightGBM. The logarithmic volatility predicted by all the models are further transformed to positive prediction. A time series grid search using only pre-test period has been used for setting regularisation strength and tree complexity. The forecast ensemble is the arithmetic mean of GJR-GARCH, Ridge, Random Forest, XGBoost, and LightGBM forecasts, each with a fixed weight of 0.20. Equal weights were chosen before out-of-sample evaluation as a simple combination across model families; they were not estimated from test-period results.

3.4 Validation and evaluation

Out-of-sample testing starts from 1 September 2023. The estimation window grows as more data become available, and the model is retrained every first trading day of each calendar quarter. At each after-close refit on trading day t , a training row dated s is included only if all five future returns in its target have become available: the position of s on the common trading-day index plus five must be no later than the position of t . Thus, the four immediately preceding trading-day rows are excluded, while a row dated five trading days earlier is eligible. Scaling, estimation, and hyperparameters tuning are based exclusively on training data. Chronological sorting prevents any possible data leak associated with random splitting of financial time series. Based on the availability of features at the end of the common period, the process results in approximately 743-745 predictions per ETF.

MAE, RMSE, bias, out-of-sample R^2 , and QLIKE has been used to evaluate forecast accuracy. The actual variance is compared with the predicted variance using QLIKE. It is calculated as the ratio of actual variance to predicted variance, minus the logarithm of that ratio, and minus one. A more accurate forecast is given when the QLIKE value is lower. A Friedman test assesses differences in model rankings across ETFs, followed by one-tailed Wilcoxon signed-rank tests with Holm correction. A 90% Hansen–Lunde–Nason Model Confidence Set was constructed using the studentized range statistic, 1,000 stationary bootstrap repetitions, and a block size of 10.

Market volatility low, normal and high states are determined using SPY's 22-day volatility terciles. The economic value of the forecasts is investigated in the context of a volatility targeting portfolio that adjusts the weight of each ETF to target 10% volatility per annum and restricts leverage to a maximum of two. At each daily rebalance, turnover is the absolute change in portfolio weight from the preceding day, with the initial position measured against zero; a transaction cost of five basis points (0.0005) per unit of turnover is deducted from that day's strategy return. The performance metrics reported include annual return and volatility, Sharpe and Sortino ratio, maximum drawdown, average turnover and 95% normal value at risk violations.

4. RESULTS

4.1 Descriptive Evidence

There are substantial differences between the ETFs in terms of how their distributions of returns look. The annualized volatility varies from 0.016 (SHY) to 0.231 (IWM). The maximum annualized logarithmic return is recorded by QQQ at 0.189, and the annualized return of TLT is negative at -0.025 during the entire period under analysis. Skewness in the distribution of all but two of the series is negative. All series have positive excess kurtosis. This information confirms the need to use heavy tailed innovations and allow for asymmetric volatility modelling.

4.2 Average Forecasting Performance

TABLE II. AVERAGE OUT-OF-SAMPLE PERFORMANCE ACROSS 12 ETFs

Model	MAE	RMSE	QLIKE	OOS R2	Bias	Mean rank
EGARCH	0.0446	0.0629	0.3817	0.1094	0.0095	1.9167
GJR-GARCH	0.0452	0.0645	0.3921	0.0498	0.0103	2.4167
GARCH	0.0468	0.0671	0.4029	-0.0025	0.0121	3.6667
EWMA	0.0462	0.0663	0.4014	0.0436	0.0087	3.6667
Ensemble	0.0412	0.0612	0.4113	0.1587	-0.0062	4.6667
Historical	0.0472	0.0686	0.4347	-0.0210	0.0070	7.1667
RandomForest	0.0422	0.0633	0.4546	0.1131	-0.0097	8.0833
LightGBM	0.0432	0.0636	0.4661	0.0936	-0.0084	9.1667
XGBoost	0.0422	0.0627	0.4580	0.1191	-0.0100	8.8333
HAR	0.0420	0.0636	0.4782	0.1129	-0.0119	10.2500
Lasso	0.0419	0.0629	0.5376	0.0932	-0.0152	10.9167
ElasticNet	0.0419	0.0629	0.5381	0.0936	-0.0150	10.7500
OLS	0.0430	0.0635	0.5371	0.0709	-0.0128	11.8333
Ridge	0.0426	0.0630	0.5382	0.0844	-0.0134	11.6667

The EGARCH is the model with the smallest mean value of QLIKE (0.3817) and the highest mean rank (1.92). The GJR-GARCH has the second rank according to the QLIKE criterion. The ensemble has the best results among the level error measures, having the minimum value of MAE (0.0412) and RMSE (0.0612). Moreover, the ensemble has the highest mean out-of-sample R-squared (0.159). The HAR produces the good MAE (0.0420), but its negative mean bias makes its QLIKE equal to 0.4782.

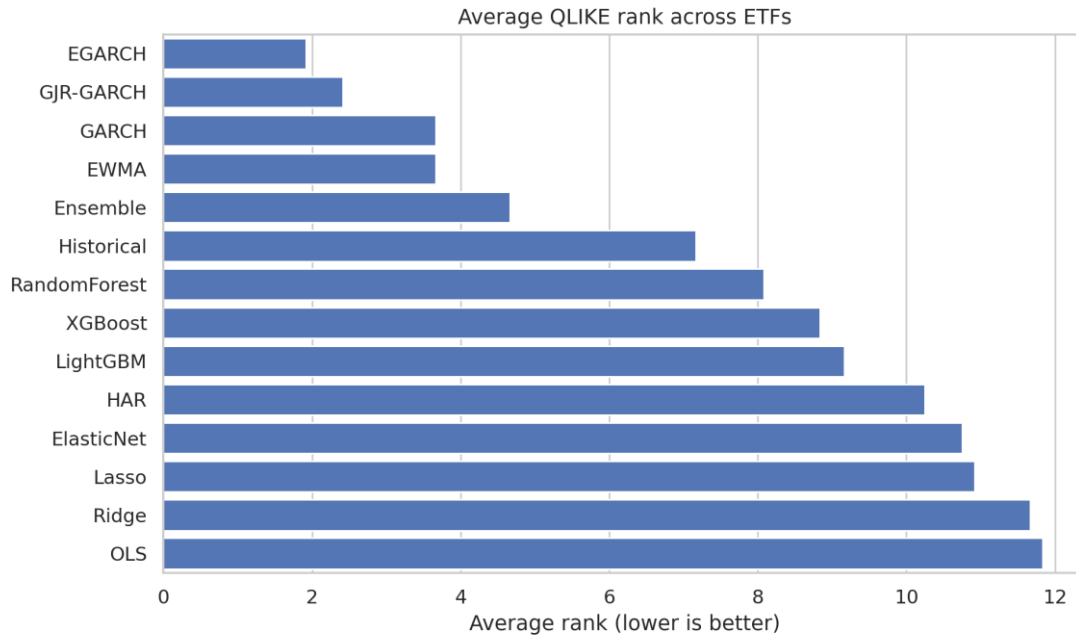


Fig.2. Average QLIKE rank across ETFs

TABLE III. BEST QLIKE MODEL FOR EACH ETF

ETF	QLIKE winner	QLIKE	MAE	RMSE	OOS R2
AGG	EWMA	0.2798	0.0147	0.0189	0.1884
DBC	GARCH	0.3495	0.0486	0.0658	0.1562
EEM	GARCH	0.4839	0.0658	0.0897	0.2014
EFA	EGARCH	0.3918	0.0466	0.0694	0.0946
GLD	EGARCH	0.3941	0.0591	0.0937	0.2713
IWM	EGARCH	0.3661	0.0657	0.0906	-0.0409
LQD	EGARCH	0.3588	0.0214	0.0287	0.0765
QQQ	GJR-GARCH	0.3986	0.0629	0.0920	0.1559
SHY	EGARCH	0.3862	0.0057	0.0073	-0.1278
SPY	Ensemble	0.3997	0.0405	0.0694	0.3258
TLT	GARCH	0.2809	0.0393	0.0495	0.1780
VNQ	EGARCH	0.3941	0.0560	0.0754	-0.0404

No single model performs best for every ETF. EGARCH gives the best results for EFA, GLD, IWM, LQD, SHY, and VNQ. GARCH performs best for DBC, EEM, and TLT. GJR-GARCH leads for QQQ, and the ensemble model leads for SPY, while EWMA performs best for AGG. Therefore, the most suitable model depends on the type of asset. The results do not support a single machine-learning model for all ETFs.

4.3 Statistical Comparisons

The Friedman test produced a statistic of 113.40 ($p = 4.02e-18$). This result shows that the models do not have equal performance rankings. After applying the Holm correction, the one-sided comparisons showed that EGARCH performed

significantly better than historical volatility, EWMA, HAR, the linear models, and the standalone tree-based models and the ensemble. However, EGARCH did not significantly outperform GARCH or GJR-GARCH at the 5% level.

The adjusted p-value was 0.0522 when EGARCH was compared with GARCH, 0.0183 with the ensemble, and 0.1331 with GJR-GARCH. The ensemble comparison is statistically significant at the 5% level; the other two are not. Comparisons based on GARCH also became less conclusive after adjusting for multiple tests, the reason partially may be the five-day forecast targets overlap. In total, though EGARCH reached the best average rank, it did not significantly outperform GARCH or GJR-GARCH.

TABLE IV. EXACT 90 PERCENT HANSEN LUNDE NASON MODEL CONFIDENCE SETS

ETF	Models retained in 90% set
AGG	EGARCH, EWMA, GARCH, GJR-GARCH, Historical
DBC	EGARCH, EWMA, ElasticNet, Ensemble, GARCH, GJR-GARCH, HAR, Historical, Lasso, OLS, RandomForest, Ridge, XGBoost
EEM	EGARCH, EWMA, ElasticNet, Ensemble, GARCH, GJR-GARCH, HAR, Historical, Lasso, OLS, Ridge
EFA	EGARCH, EWMA, Ensemble, GARCH, GJR-GARCH, HAR, Historical
GLD	EGARCH, EWMA, Ensemble, GARCH, GJR-GARCH, Historical, OLS
IWM	EGARCH, EWMA, ElasticNet, Ensemble, GARCH, GJR-GARCH, HAR, Lasso, LightGBM, RandomForest
LQD	EGARCH, EWMA, Ensemble, GARCH, GJR-GARCH, HAR, Historical, LightGBM, RandomForest
QQQ	EGARCH, EWMA, ElasticNet, Ensemble, GARCH, GJR-GARCH, Historical, Lasso, LightGBM, OLS, RandomForest, Ridge
SHY	EGARCH, EWMA
SPY	EGARCH, ElasticNet, Ensemble, GJR-GARCH, Lasso, OLS, RandomForest, Ridge, XGBoost
TLT	EGARCH, EWMA, ElasticNet, Ensemble, GARCH, GJR-GARCH, HAR, Historical, Lasso, LightGBM, OLS, Ridge, XGBoost
VNQ	EGARCH, EWMA, ElasticNet, Ensemble, GARCH, GJR-GARCH, HAR, Historical, Lasso, LightGBM

The Model Confidence Sets reveal a more conservative view on the rankings of models. EGARCH is included in the confidence sets of all 12 ETFs. EWMA and GJR-GARCH are each included in the sets for 11 ETFs, while GARCH and the ensemble are each included for 10 ETFs. SHY gives rise to the most selective confidence set, containing only EGARCH and EWMA. At the same time, there are several ETFs that have relatively large confidence sets. Thus, the inclusion into the top 3 of models' ranks does not imply their uniqueness when the evaluation sample is finite and serially dependent.

4.4 Regime Dependence

TABLE V. BEST QLIKE MODEL BY SPY VOLATILITY REGIME

Regime	Model	QLIKE	MAE	RMSE
High	EGARCH	0.4208	0.0517	0.0833
Low	GARCH	0.4186	0.0422	0.0649
Normal	EGARCH	0.3040	0.0399	0.0562

The leading model is conditional upon market conditions. In low-volatility regimes, GARCH proves to perform better than other models; EGARCH is the best in normal conditions, and EGARCH also leads in high-volatility periods. The good performance of EGARCH during turbulence may reflect its asymmetric volatility specification.

4.5 Explainability

Mean absolute SHAP values of features, calculated across tree models and ETFs, identify those that move the fitted predictions the most. Trend displacement on a medium horizon, cross-asset risk conditions, and downside volatility indicators prove to be some of the leaders. Individual return lags play a comparatively minor role in predicting returns. Mean absolute SHAP values describe feature importance and do not characterize any kind of causal influence. Correlated features might share responsibility.

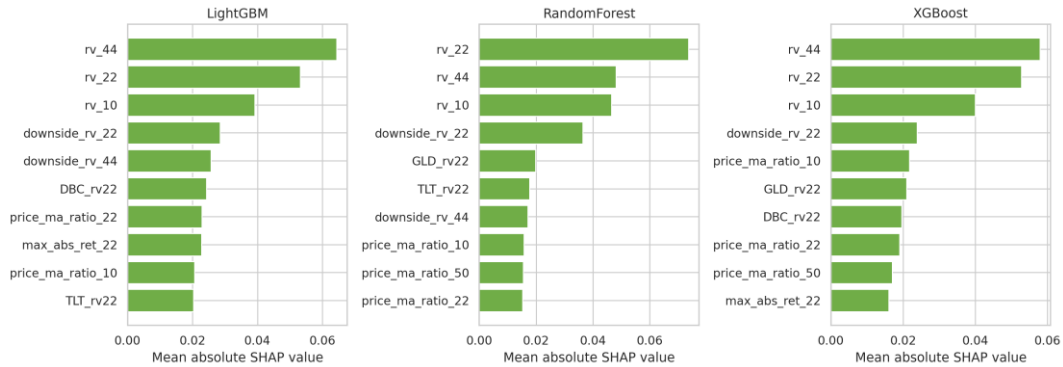


Fig. 3. Leading predictors in the tree models

4.6 Economic Evaluation

TABLE VI. MEAN VOLATILITY-TARGETING OUTCOMES ACROSS ETFs

Model	Return	Volatility	Sharpe	Sortino	Max drawdown	Turnover	VaR violations
HAR	0.090	0.112	0.895	1.512	-0.117	0.032	0.072
Ensemble	0.083	0.105	0.874	1.482	-0.111	0.036	0.065
EWMA	0.077	0.098	0.877	1.481	-0.104	0.022	0.053
Lasso	0.088	0.113	0.867	1.469	-0.118	0.047	0.078
GJR-GARCH	0.074	0.096	0.868	1.459	-0.105	0.030	0.047
ElasticNet	0.088	0.113	0.865	1.467	-0.118	0.049	0.077
XGBoost	0.085	0.109	0.867	1.475	-0.117	0.051	0.070
EGARCH	0.074	0.096	0.862	1.450	-0.105	0.033	0.045
Historical	0.077	0.101	0.860	1.447	-0.109	0.024	0.058
Random Forest	0.084	0.110	0.854	1.444	-0.116	0.041	0.069

HAR has the largest average Sharpe ratio in the baseline volatility targeting (0.895), followed by EWMA (0.877) and the ensemble (0.874). Despite EGARCH's dominance in the QLIKE rankings, it still achieves a Sharpe ratio of 0.862. Furthermore, HAR exhibits higher realized volatility and has a 7.2% VaR violation rate. Thus, the higher Sharpe ratio does not indicate improved tail-risk forecasting ability of this method. EGARCH and GJR-GARCH hold their VaR violation rates at or below about 5%. The economic ranking depends on the realized return and position sizes generated by the forecasts; therefore, the results can be considered descriptive and cannot be taken as evidence of the superior ability of a specific strategy.

Actual and predicted five-day volatility using EGARCH

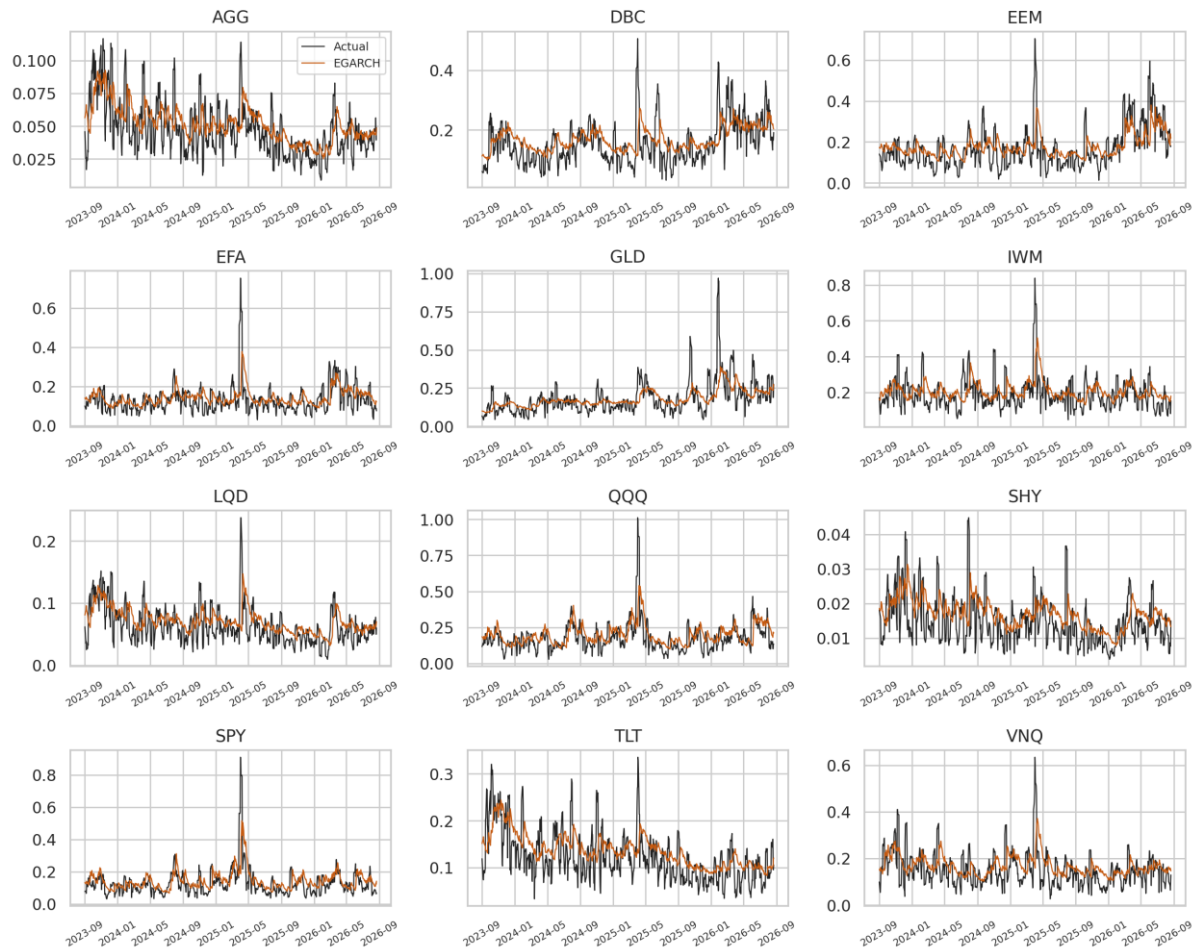


Fig.4. Actual and predicted volatility from the overall QLIKE winner

5. DISCUSSION

As for the research questions, the answers depend on the criteria used. Under the main QLIKE criterion, EGARCH provides the most accurate forecasts and thus answers RQ1. However, the ensemble proves more accurate under level error, and HAR demonstrates the highest baseline Sharpe ratio. As for RQ2, the significant rank differences established by the rank tests corroborate the existence of genuine differences between the models. At the same time, the Model Confidence Set holds several statistically equivalent alternatives for each ETF. RQ3 is supported because the best-performing model changes across ETFs, asset classes, and volatility regimes. For RQ4, the SHAP results show that volatility, downside risk, trend movements, and cross-market risk are important predictors. However, these relationships should not be interpreted as causal. RQ5 receives a mixed answer because a model with better statistical accuracy does not always produce better economic results.

The findings are consistent with the model structure for EGARCH. First, the model is specified using log variance that prevents the negative forecasts. Second, there is persistence and asymmetry built-in into the specification. The restrictions of the models might be beneficial with the sample size of about 1,700 observations before the test period for each ETF. The tree models operate a larger set of predictors and learn many nonlinearities using a smaller effective sample that changes over time. The negative average bias explains their performance under MAE and RMSE, yet the inferior performance under QLIKE.

As for the ensemble, combining predictions made by different models can reduce the model-specific errors and increase the accuracy of predictions in MAE and RMSE. An equal weight combination can inherit the downward bias existing in the machine learning models that are part of it. Future research can try estimating the forecast combination weights based

on QLIKE estimated before the test period or using the model confidence set. It is important to estimate all weights in advance.

From a practical perspective, the choice of model should depend on how the forecast will be used. EGARCH is suitable when underestimating volatility could lead to serious risk and QLIKE is the main evaluation measure. The ensemble model is useful when reducing level errors and improving volatility-scaled returns are the main priorities. EWMA remains a reasonable choice for low-volatility fixed-income assets because it generally produces lower portfolio turnover.

6. ROBUSTNESS, LIMITATIONS, AND FUTURE RESEARCH

TABLE VII. ROBUSTNESS WINNERS ACROSS EVALUATION SAMPLES (RMSE POOLED ACROSS ETF-DAY FORECASTS)

Sample	Model	QLIKE	MAE	RMSE
From 2024-01	EGARCH	0.3906	0.0450	0.0706
From 2025-01	EGARCH	0.4168	0.0472	0.0780
Full	EGARCH	0.3817	0.0446	0.0691
Non-overlapping 5-day	EGARCH	0.3680	0.0442	0.0665

TABLE VIII. BEST QLIKE MODEL BY ASSET EXPOSURE

AssetClass	Model	QLIKE
Aggregate bonds	EWMA	0.2798
Commodities	GARCH	0.3495
Corporate bonds	EGARCH	0.3588
Developed-market equity	EGARCH	0.3918
Emerging-market equity	GARCH	0.4839
Gold	EGARCH	0.3941
Long Treasury bonds	GARCH	0.2809
Real estate	EGARCH	0.3941
Short Treasury bonds	EGARCH	0.3862
US growth equity	GJR-GARCH	0.3986
US large-cap equity	Ensemble	0.3997
US small-cap equity	EGARCH	0.3661

EGARCH remains the best overall model based on QLIKE when the evaluation starts in either January 2024 or January 2025. It also performs best when every fifth observation is selected to create non-overlapping five-day forecasts. However, the results vary across asset classes. EGARCH performs best for developed-market equities, gold, real estate, short-term Treasuries, and small-cap equities. GARCH performs best for commodities, emerging-market equities, and long-term Treasuries. EGARCH performs best for corporate bonds, GJR-GARCH for US growth equities, and the ensemble for US large-cap equities. EWMA performs best for aggregate bonds.

TABLE IX. ECONOMIC SENSITIVITY WINNERS

Cost bps	Leverage cap	Best model	Mean Sharpe
0	1.0	HAR	0.967
0	2.0	HAR	0.930
5	1.0	HAR	0.943

Cost bps	Leverage cap	Best model	Mean Sharpe
5	2.0	HAR	0.895
10	1.0	HAR	0.919
10	2.0	HAR	0.860

HAR keeps the highest Sharpe ratio on average – for zero, five, and ten basis points of transaction costs and for leverage caps of one and two. Its Sharpe ratio drops from 0.967 without transaction costs and unit leverage cap to 0.860 in case of ten basis points transaction costs and two leverage cap. The winner remains the same, but the magnitude of its economic edge is evidently dependent on the implemented assumptions. Some limitations of this study must be mentioned. First, the study relies on close-to-close returns because high–low prices were excluded and intraday data were unavailable. As a result, the volatility measure may contain some estimation error and differs from the intraday realised variance typically used in HAR studies. Second, the five-day forecast periods overlap. The non-overlapping robustness check reduces this dependence, but may not eliminate its effects in a finite sample. Third, the ten-year sample includes several unusual market periods, such as the COVID-19 pandemic, high inflation, and major interest-rate changes. Therefore, the results may not be the same under more stable market conditions. Fourth, the models rely only on ETF prices and trading volumes. Future research could include macroeconomic variables and option-implied measures to improve forecasting accuracy.

Quarterly refitting is another limitation. Although it allows the models to be updated regularly while keeping the computational workload manageable, monthly or more frequent refitting could produce different results. The economic analysis also uses a simple normal VaR approach that does not fully account for financing costs, taxes, bid–ask spreads, or market impact.

In addition, SHAP may distribute importance unevenly among predictors containing similar information. Comparing many models across several ETFs also increases the risk of data snooping. Training-only tuning, fixed evaluation rules, the Holm correction, and the Model Confidence Set help reduce this risk, but they cannot eliminate it entirely.

Future research could incorporate intraday realised volatility, the VIX, and yield-curve indicators. It could also examine one-day and 22-day forecast horizons separately and test alternative EWMA decay factors and model-refitting frequencies. Regime-switching and quantile-based models may offer further insights, while nested validation could improve the way forecasts are combined. Finally, a preregistered study using data collected after August 2026 could determine whether the findings remain reliable over time.

7. CONCLUSION

This study evaluated 14 volatility-forecasting models across 12 diversified US ETFs using a consistent out-of-sample framework designed to prevent data leakage. EGARCH delivered the strongest overall performance based on the QLIKE measure, reducing the loss by 12.2% compared with historical volatility and by 5.3% compared with GARCH. It also remained the best-performing model when later evaluation periods and non-overlapping forecast targets were considered. After the Holm adjustment, EGARCH achieved a significantly better average rank than the tree-based models, linear models, HAR, historical volatility, EWMA, and the ensemble. However, it did not significantly outperform GARCH or GJR-GARCH at the 5% level. The ensemble produced the lowest MAE and RMSE and the highest out-of-sample R2. HAR achieved the highest volatility-targeting Sharpe ratio, although its VaR calibration was weaker. The Model Confidence Set selected EGARCH for all ETFs, while several other models also performed well for most assets. These results suggest that asymmetric econometric models remain strong competitors. Combining forecasts may help reduce prediction errors, while HAR can be particularly useful for position-sizing decisions. Therefore, the most suitable model depends on the asset, volatility regime, evaluation measure, and purpose of the forecast.

Conflicts of Interest

The authors declare no conflict of interest.

Funding

This research received no external funding.

Acknowledgment

None

References

- [1] R. F. Engle, “Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation,” *Econometrica*, vol. 50, no. 4, pp. 987–1007, 1982, doi: 10.2307/1912773.
- [2] T. Bollerslev, “Generalized autoregressive conditional heteroskedasticity,” *Journal of Econometrics*, vol. 31, no. 3, pp. 307–327, 1986, doi: 10.1016/0304-4076(86)90063-1.
- [3] L. R. Glosten, R. Jagannathan, and D. E. Runkle, “On the relation between the expected value and the volatility of the nominal excess return on stocks,” *The Journal of Finance*, vol. 48, no. 5, pp. 1779–1801, 1993, doi: 10.1111/j.1540-6261.1993.tb05128.x.
- [4] D. B. Nelson, “Conditional heteroskedasticity in asset returns: A new approach,” *Econometrica*, vol. 59, no. 2, pp. 347–370, 1991, doi: 10.2307/2938260.
- [5] F. Corsi, “A simple approximate long-memory model of realized volatility,” *Journal of Financial Econometrics*, vol. 7, no. 2, pp. 174–196, 2009, doi: 10.1093/jfinec/nbp001.
- [6] L. Breiman, “Random forests,” *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001, doi: 10.1023/A:1010933404324.
- [7] T. Chen and C. Guestrin, “XGBoost: A scalable tree boosting system,” in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Francisco, CA, USA, 2016, pp. 785–794, doi: 10.1145/2939672.2939785.
- [8] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T.-Y. Liu, “LightGBM: A highly efficient gradient boosting decision tree,” in *Advances in Neural Information Processing Systems 30*, Long Beach, CA, USA, 2017, pp. 3146–3154.
- [9] P. R. Hansen and A. Lunde, “A forecast comparison of volatility models: Does anything beat a GARCH(1,1)?” *Journal of Applied Econometrics*, vol. 20, no. 7, pp. 873–889, 2005, doi: 10.1002/jae.800.
- [10] A. E. Hoerl and R. W. Kennard, “Ridge regression: Biased estimation for nonorthogonal problems,” *Technometrics*, vol. 12, no. 1, pp. 55–67, 1970, doi: 10.1080/00401706.1970.10488634.
- [11] R. Tibshirani, “Regression shrinkage and selection via the lasso,” *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 58, no. 1, pp. 267–288, 1996, doi: 10.1111/j.2517-6161.1996.tb02080.x.
- [12] H. Zou and T. Hastie, “Regularization and variable selection via the elastic net,” *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 67, no. 2, pp. 301–320, 2005, doi: 10.1111/j.1467-9868.2005.00503.x.
- [13] A. J. Patton, “Volatility forecast comparison using imperfect volatility proxies,” *Journal of Econometrics*, vol. 160, no. 1, pp. 246–256, 2011, doi: 10.1016/j.jeconom.2010.03.034.
- [14] F. X. Diebold and R. S. Mariano, “Comparing predictive accuracy,” *Journal of Business & Economic Statistics*, vol. 13, no. 3, pp. 253–263, 1995, doi: 10.1080/07350015.1995.10524599.
- [15] P. R. Hansen, A. Lunde, and J. M. Nason, “The model confidence set,” *Econometrica*, vol. 79, no. 2, pp. 453–497, 2011, doi: 10.3982/ECTA5771.
- [16] C. Bergmeir, R. J. Hyndman, and B. Koo, “A note on the validity of cross-validation for evaluating autoregressive time series prediction,” *Computational Statistics & Data Analysis*, vol. 120, pp. 70–83, 2018, doi: 10.1016/j.csda.2017.11.003.