

Volatility Analysis of Cryptocurrency Markets: A Comparative Study of EGARCH, LSTM and Hybrid Forecasting Models for Bitcoin and Ethereum

Keith Reuben Muyungi¹; Hu Junjuan²

School of Science, Zhejiang University of Science and Technology, Hangzhou, China

Publication Date: 2026/09/18

Abstract: This paper compares econometric and deep learning approaches to one-day-ahead volatility forecasting for Bitcoin (BTC/USD) and Ethereum (ETH/USD) over the period from January 2016 to December 2025. Seven models are evaluated: a symmetric GARCH(1,1) with Student-t innovations, an asymmetric EGARCH(1,1) with Student-t innovations, a Long Short-Term Memory neural network, two hybrid LSTM specifications, a constant unconditional-variance benchmark, and a naive random walk. Hybrid-A augments the LSTM input with the EGARCH conditional variance forecast. Hybrid-B applies the LSTM to the standardised residuals of the fitted EGARCH model. To guard against the well-documented sensitivity of recurrent neural networks to random initialisation, the LSTM-based models are estimated under ten different random seeds and the median forecast is reported. Squared daily log returns serve as the primary volatility proxy, with the Parkinson and Garman-Klass range-based estimators used as robustness benchmarks. Forecasts are produced through an expanding-window rolling scheme with monthly refits over the test period from January 2023 to December 2025, yielding 1,096 out-of-sample forecasts per model and asset. Statistical inference is based on the Diebold-Mariano test with the Harvey-Leybourne-Newbold correction and the Model Confidence Set procedure of Hansen, Lunde, and Nason (2011). The principal finding is that the relative ranking of cryptocurrency volatility forecasts is highly sensitive to the choice of loss function but stable across alternative volatility proxies. Under MSE, the LSTM achieves the lowest error and is the only model retained in the 90% Model Confidence Set for Ethereum, while several models survive jointly for Bitcoin. Under QLIKE, which is robust to noise in the squared-return proxy and penalises variance underestimation, both GARCH and EGARCH survive the Model Confidence Set jointly with the LSTM and Hybrid-B for Ethereum, and GARCH retains a place for Bitcoin. The LSTM-based models beat the constant unconditional-variance benchmark by 37% to 72% across proxies, confirming that the LSTM captures genuine volatility dynamics rather than merely predicting the mean. The variance-as-feature hybrid is dominated by the pure LSTM and exhibits substantially higher seed-to-seed variability than the other deep learning specifications, while the residual-based hybrid effectively collapses to the pure LSTM. Sub-period analysis confirms that these patterns are stable across the 2023, 2024, and post-spot-ETF 2025 windows. The findings indicate that the apparent superiority of LSTM models in cryptocurrency volatility forecasting is partly an artefact of evaluation under squared error and that simple parametric models retain a useful role under loss-consistent inference.

Keywords: Cryptocurrency Volatility; EGARCH; Long Short-Term Memory; Hybrid Forecasting.

How to Cite: Keith Reuben Muyungi; Hu Junjuan (2026) Volatility Analysis of Cryptocurrency Markets: A Comparative Study of EGARCH, LSTM and Hybrid Forecasting Models for Bitcoin and Ethereum. *International Journal of Innovative Science and Research Technology*, 11(8), 3901-3921. <https://doi.org/10.38124/ijisrt/26aug1506>

I. INTRODUCTION

The market capitalisation of digital assets has increased from a few hundred million dollars in 2013 to several trillion dollars by 2025, with Bitcoin and Ethereum together accounting for the bulk of trading activity. As cryptocurrencies have moved from the margins of speculative finance to the portfolios of pension funds, sovereign wealth investors, and listed exchange-traded products, the question of how to model and forecast their volatility has become a

pressing concern for academics, regulators, and practitioners. Volatility forecasts feed directly into value-at-risk computations, option pricing, position sizing, and the design of margin requirements; their quality therefore has consequences that extend beyond statistical accuracy and into the stability of the broader financial system (Babalos et al., 2025; Whaley & Hornback, 2025).

Cryptocurrencies pose unique challenges for volatility modelling and unlike traditional currencies or equities, they

lack a fundamental anchor, trade twenty-four hours a day across fragmented venues, and exhibit episodes of extreme price dislocation associated with hacks, regulatory shocks, and speculative bubbles hence the resulting return distributions are heavy-tailed, asymmetric, and prone to abrupt regime shifts. Standard parametric volatility models, most notably the Generalised Autoregressive Conditional Heteroskedasticity (GARCH) framework of Bollerslev (1986), often struggle to capture this behaviour, with several authors reporting that key parameter estimates are unstable or fail to satisfy the stationarity restrictions imposed by the model (Caporale & Zekokh, 2019; Trucíos, 2019).

Two parallel literatures have emerged in response. The first refines the parametric framework through asymmetric specifications such as EGARCH (Nelson, 1991), regime-switching extensions (Ardia et al., 2019; Caporale & Zekokh, 2019), and long-memory variants (Kaya Soylu et al., 2020). The second turns to machine learning techniques, in particular Long Short-Term Memory (LSTM) recurrent neural networks (Hochreiter & Schmidhuber, 1997), which can in principle approximate arbitrary nonlinear dependencies in sequential data without imposing the distributional assumptions that constrain GARCH-type models. A third strand combines the two by constructing hybrid models in which a GARCH process supplies a volatility signal to a neural network, or in which a neural network operates on the residuals of a fitted GARCH model (Kristjanpoller & Minutolo, 2018; Seo & Kim, 2020; Zahid et al., 2022; García-Medina & Aguayo-Moreno, 2024).

Despite a growing volume of empirical work, the comparative performance of these three classes of models remains unsettled. Some studies report that LSTM models substantially outperform GARCH-type benchmarks for Bitcoin volatility (Pratas et al., 2023; Phan Huy et al., 2025), while others find that simple parametric models are competitive or even superior once forecasts are evaluated rigorously over multiple horizons and loss functions (Köchling et al., 2020; Bergsli et al., 2022; Dudek et al., 2024). The literature on hybrid models is similarly mixed, with evidence that EGARCH-LSTM combinations improve in-sample fit but do not always translate into out-of-sample gains (Amirshahi & Lahmiri, 2023; Araya et al., 2024). The lack of consensus reflects differences in sample period, forecasting horizon, evaluation procedure, and the choice of cryptocurrency, which together make direct comparisons across studies difficult.

This paper firstly provides a unified comparison of five models, namely GARCH(1,1), EGARCH(1,1), pure LSTM, Hybrid-A, and Hybrid-B, applied consistently to Bitcoin and Ethereum over the ten-year window from January 2016 to December 2025. The sample spans the 2017 bull run, the 2018 crash, the COVID-19 shock, the 2021 expansion, the 2022 deleveraging episode following the collapse of Terra-Luna and FTX, and the post-spot-ETF environment of 2024 and 2025, providing a rich variety of volatility regimes against which to benchmark the models. Secondly, an explicit comparison between two distinct hybrid architectures is introduced. Hybrid-A augments the LSTM input with the

one-step-ahead conditional variance produced by the EGARCH model, following the dominant approach in the literature. Hybrid-B applies the LSTM to the standardised residuals of the EGARCH model and recombines the two components multiplicatively, in the spirit of Araya et al. (2024) and Zhao et al. (2025). The aim is to let the neural network capture residual nonlinearity that the parametric model misses, rather than duplicating the information already contained in the EGARCH variance.

Third, we two methodological concerns that recur in the deep learning literature. The performance of recurrent neural networks is known to depend on the random initialisation of weights and the order in which training samples are presented (Reimers & Gurevych, 2017). Reporting results from a single training run, as is common in this literature, risks overstating the precision of the conclusions. Each LSTM-based model is therefore estimated under ten different random seeds, and the median forecast across seeds is used in the headline evaluation. The seed-to-seed distribution of metrics is reported separately, providing both robustness and a window onto the stability properties of the alternative specifications. The suitability of the EGARCH baseline is also assessed through Ljung-Box and ARCH-LM tests on the standardised residuals of the fitted model, ensuring that the parametric component is well specified before being used as the foundation of the hybrids.

Fourth, we apply a comprehensive evaluation framework that combines five accuracy metrics with the Diebold-Mariano test (Diebold & Mariano, 1995) using the correction of Harvey, Leybourne, and Newbold (1997) and the Model Confidence Set procedure of Hansen, Lunde, and Nason (2011). This dual approach addresses the well-known concern that point comparisons of forecast accuracy can be misleading when the loss function is sensitive to outliers, and it is particularly important for cryptocurrencies whose squared returns are highly skewed (Patton, 2011; Liu et al., 2015). Sub-period analysis across 2023, 2024, and 2025 is used to verify that the rankings are stable across regimes.

The principal findings can be summarised as follows. First, the relative ranking of cryptocurrency volatility forecasts is highly sensitive to the choice of loss function. Under MSE, the LSTM produces the most accurate point forecasts of one-day-ahead variance for both Bitcoin and Ethereum, and is the only model retained in the 90% Model Confidence Set for Ethereum. Under the QLIKE loss, which is robust to noise in the squared-return proxy and penalises variance underestimation more heavily, the parametric models survive jointly with the LSTM and Hybrid-B in the Model Confidence Set: GARCH(1,1)-t and EGARCH(1,1)-t are both retained for Ethereum, and GARCH(1,1)-t is retained for Bitcoin. Closer inspection of the forecast paths reveals why the two metrics diverge: the LSTM tends to produce variance forecasts close to the unconditional mean and thereby achieves low average squared error, while the parametric models respond more aggressively to volatility regimes and pay a higher MSE cost when their predictions overshoot but a comparable QLIKE cost when realised variance is itself elevated.

Second, the symmetric GARCH(1,1)-t baseline is competitive with EGARCH and is occasionally preferred. The asymmetric leverage parameter γ in the EGARCH specification is statistically weak for Bitcoin and statistically insignificant for Ethereum, suggesting that the asymmetric volatility response that justifies the use of EGARCH in equity markets is muted in this cryptocurrency sample. EGARCH still attains a slightly lower AIC and BIC than GARCH, but the two models are statistically indistinguishable in out-of-sample forecasting under both MSE and QLIKE losses. Third, the variance-as-feature hybrid (Hybrid-A) is consistently dominated by the pure LSTM and exhibits substantially higher seed-to-seed variability, with the MSE for Ethereum ranging from 664 to 1,418 across the ten random seeds. The residual-based hybrid (Hybrid-B), by contrast, is statistically indistinguishable from the pure LSTM in every metric and across both assets.

These findings carry practical implications. For applications in which point accuracy of variance forecasts is the primary concern, such as parameter inputs to portfolio optimisation, the LSTM offers measurable improvements over the parametric benchmarks. For risk management applications, where downside protection requires that variance not be systematically underestimated, the parametric models retain a useful role under QLIKE-based evaluation. The results also caution against the assumption that hybrid models automatically outperform their components: when the parametric model is misspecified relative to the data-generating process, feeding its variance signal as an input feature can degrade rather than improve the LSTM's forecast and introduce material instability across training runs.

The remainder of the paper proceeds as follows. Section 2 reviews the relevant literature on cryptocurrency volatility forecasting. Section 3 describes the data and reports descriptive statistics together with stylised-fact diagnostics and EGARCH residual checks. Section 4 develops the EGARCH, GARCH, LSTM, and hybrid models and explains the forecasting and evaluation procedures. Section 5 presents the empirical results, and Section 6 discusses their implications. Section 7 concludes.

II. LITERATURE REVIEW

➤ *Parametric Volatility Models for Cryptocurrencies*

The parametric tradition in cryptocurrency volatility modelling is rooted in the GARCH framework of Bollerslev (1986). Katsiampa (2017) was among the first to apply a wide class of GARCH-type models to Bitcoin returns and concluded that the AR(1)-CGARCH specification provided the best in-sample fit. Subsequent work has extended this comparison to broader sets of cryptocurrencies and richer model classes. Fakhfekh and Jeribi (2020) examine the volatility dynamics of sixteen cryptocurrencies and find that asymmetric specifications, including EGARCH and TGARCH, consistently outperform the symmetric GARCH(1,1) baseline. Cheikh et al. (2020) document an asymmetric volatility response in cryptocurrency markets that runs in the opposite direction to that observed in equity markets, with positive shocks generating a larger volatility

increase than negative shocks of equal magnitude. Katsiampa (2019) and Ngunyi et al. (2019) confirm this pattern across additional cryptocurrencies. Conrad et al. (2018) decompose Bitcoin volatility into long-run and short-run components and report that macroeconomic and financial market factors influence the long-run component.

Several authors have raised concerns about the stability of GARCH-type estimates for cryptocurrencies. Caporale and Zekokh (2019) and Ardia et al. (2019) respond by introducing Markov-switching GARCH models that allow the volatility process to switch between regimes; both studies report meaningful improvements in forecast accuracy and value-at-risk performance relative to single-regime alternatives. Kaya Soylu et al. (2020) investigate long memory in cryptocurrency volatility and find significant fractional integration parameters, motivating the use of FIGARCH and HYGARCH models. Trucíos (2019) and Trucíos and Taylor (2023) develop robust estimation procedures that down-weight outliers and report substantial improvements in value-at-risk and expected-shortfall forecasts. More recent contributions by Mohamed et al. (2025) and Sözen (2025) compare a wide spectrum of GARCH-family models on Bitcoin and conclude that asymmetric specifications such as EGARCH continue to provide a competitive baseline. Amendola and Sensini (2020) and Maciel (2021) reach similar conclusions on broader cryptocurrency panels. Fiszeder et al. (2024) propose range-based GARCH estimators that exploit intraday high-low information and deliver further accuracy gains.

Beyond the GARCH family, the heterogeneous autoregressive (HAR) model of Corsi (2009) has been applied to cryptocurrencies with mixed results. Bergsli et al. (2022) report that HAR-type models based on realised variance outperform GARCH-type models for Bitcoin, while Dudek et al. (2024) find that simple HAR and ridge-regression specifications are competitive with much more complex machine learning methods for Bitcoin, Ethereum, Litecoin, and Monero. Aras et al. (2025) extend the parametric framework by incorporating uncertainty and investor attention indices and document that they improve the explanatory power of GARCH models for Bitcoin.

➤ *Machine Learning and Deep Learning Approaches*

The application of machine learning to cryptocurrency volatility has expanded rapidly in the past five years. Early studies focused on shallow architectures such as multilayer perceptrons and support vector regression. D'Amato et al. (2022) report that deep learning models, including nonlinear autoregressive networks and Jordan networks, outperform linear benchmarks for Bitcoin, Ethereum, and Ripple volatility.

More recent work has shifted to recurrent architectures, in particular the LSTM network of Hochreiter and Schmidhuber (1997), which is designed to capture both short-range and long-range temporal dependencies (Staudemeyer & Morris, 2019). Pratas et al. (2023) compare LSTM, recurrent neural networks, multilayer perceptrons, and GARCH on Bitcoin and conclude that the LSTM achieves the

best out-of-sample accuracy. Phan Huy et al. (2025) extend this comparison to a broader set of cryptocurrencies and confirm that deep learning methods outperform GARCH-type benchmarks under MSE and MAE losses. Huang et al. (2024) further document that ensemble machine learning models improve forecast accuracy for Bitcoin volatility, while Brauneis and Sahiner (2024) show that incorporating sentiment indicators yields additional gains. In the value-at-risk literature, Petneházi (2021), Chronopoulos et al. (2024), and Buczynski and Chlebus (2024) demonstrate that quantile-regression neural networks and GARCH-augmented neural networks compete favourably with parametric value-at-risk models. Ormaniec et al. (2024) provide a careful comparison of LSTM and GARCH for value-at-risk and find that the relative ranking depends sensitively on the data and evaluation procedure.

A counter-current of evidence cautions against overstating the deep learning advantage. Köchling et al. (2020) find that GARCH-type models remain difficult to beat once forecasts are evaluated under heteroskedasticity-robust loss functions. Rodikov and Antulov-Fantulin (2022) report that LSTM models do not consistently outperform GARCH for Bitcoin and Ethereum and argue that the apparent advantage in some studies arises from an inappropriate choice of evaluation metric. Dudek et al. (2024) provide the most extensive comparison to date, applying twelve forecasting methods to four cryptocurrencies, and conclude that no single model dominates across all assets and loss functions and that simple linear models perform competitively with LSTM and random forests. A related concern, raised by Reimers and Gurevych (2017) in the context of natural language processing and re-emphasised by Henderson et al. (2018) for reinforcement learning, is that a single training run of a recurrent network can produce metrics that vary substantially across random seeds; the present paper takes this concern seriously by reporting seed-distribution statistics throughout.

➤ *Hybrid GARCH-Neural Network Models*

The third strand of the literature seeks to combine the structural strengths of GARCH with the flexibility of neural networks. Kristjanpoller and Minutolo (2018) introduce a hybrid framework that integrates GARCH, technical analysis indicators, and an artificial neural network for Bitcoin volatility, reporting improvements over either component used in isolation. Seo and Kim (2020) propose hybrid neural network models that incorporate GARCH-type variance forecasts as additional inputs and document substantial accuracy gains for Bitcoin. Aras (2021) develops stacking hybrid GARCH models for Bitcoin volatility forecasting that combine multiple GARCH-type variants with machine learning meta-learners. Zahid et al. (2022) extend this idea to a broader family of recurrent architectures, including bidirectional LSTM and gated recurrent units, and find that hybrid GARCH-deep-learning models dominate single-component benchmarks for Bitcoin.

Amirshahi and Lahmiri (2023) provide a large-scale comparison of hybrid GARCH-deep-learning models across twenty-seven cryptocurrencies and conclude that hybrids generally improve forecast accuracy relative to individual

GARCH or deep learning models, but that the magnitude of the improvement varies across assets. García-Medina and Aguayo-Moreno (2024) implement an LSTM-GARCH hybrid for cryptocurrency portfolio variance forecasting and report gains over both components separately. Araya et al. (2024) propose a hybrid approach in which the LSTM is trained on the standardised residuals of a GARCH model, rather than on the raw return series, and document modest but consistent improvements. Most recently, Zhao et al. (2025) introduce a GARCH-GRU hybrid that integrates GARCH-implied variance into a gated recurrent unit and demonstrate superior out-of-sample performance.

Despite this growing body of work, the comparative evaluation of alternative hybrid architectures has received less attention. Most studies adopt a single hybrid specification, typically the variance-as-feature approach, without contrasting it against alternative constructions. The present paper addresses this gap by implementing two contrasting hybrids on the same data and evaluating them under a common framework that explicitly accounts for training-run variability.

➤ *Forecast Evaluation Methodology*

A recurring methodological theme in this literature is the choice of evaluation metric. Patton (2011) demonstrates that the use of an imperfect proxy for latent volatility can reverse the ranking of competing forecasts under symmetric loss functions, and identifies the QLIKE loss as one of the small class of losses that yield consistent rankings under such proxies. Hansen and Lunde (2006) provide complementary results for the MSE loss when the proxy is unbiased. Liu et al. (2015) confirm that five-minute realised variance is a high-quality benchmark, but that lower-frequency proxies remain admissible when intraday data are unavailable. Andersen and Bollerslev (1998) and Andersen et al. (2003) establish the theoretical foundations for using realised volatility measures as ex-post benchmarks.

Hansen, Lunde, and Nason (2011) develop the Model Confidence Set procedure, which produces, at a given confidence level, a set of models that are not statistically distinguishable from the best. The procedure has become standard in the volatility forecasting literature (Catania & Grassi, 2022; Dudek et al., 2024) and is particularly useful when, as here, several models are compared simultaneously. Bernardi and Catania (2018) describe a widely used implementation. The Diebold and Mariano (1995) test with the small-sample correction of Harvey, Leybourne, and Newbold (1997) provides a complementary tool for pairwise comparisons. Bates and Granger (1969) and Granger and Ramanathan (1984) develop the foundational framework for combining forecasts, while Quaedy (2021) and Politis and Romano (1994) discuss multi-horizon and bootstrap-based extensions, respectively. The present paper follows this evaluation tradition by combining MSE, MAE, RMSE, QLIKE, Mincer-Zarnowitz R^2 , Diebold-Mariano tests, and the Model Confidence Set procedure. Walther et al. (2019) and Alves et al. (2025) further document that exogenous drivers and structural spillovers between cryptocurrencies and traditional assets matter for volatility forecasting,

although they are beyond the scope of the present analysis. Ayadi et al. (2025) document that crisis episodes such as COVID-19 and the Russia-Ukraine war substantially alter the volatility properties of major cryptocurrencies.

to improve the numerical stability of the maximum-likelihood estimation of the GARCH and EGARCH models:

$$r_t = 100 \times \ln(P_t / P_{t-1}) \quad (1)$$

where P_t denotes the closing price on day t . The squared daily log return r_t^2 serves as the primary volatility proxy throughout the analysis, following standard practice when intraday data are unavailable (Andersen & Bollerslev, 1998; Patton, 2011). Although the squared return is unbiased for the latent variance, it is also extremely noisy, a property that motivates the use of the QLIKE loss function alongside MSE in the evaluation framework. The Parkinson and Garman-Klass range-based estimators were also computed from the daily OHLC data and yielded qualitatively similar stylised facts; their use as alternative forecast targets is discussed in Section 6 as a direction for future work.

➤ Descriptive Statistics

Table 1 reports descriptive statistics for the daily log returns of Bitcoin and Ethereum over the full sample. Both assets exhibit positive average returns over the decade, with Ethereum delivering a slightly higher mean daily return of 0.220% compared with 0.145% for Bitcoin. The standard deviation of daily returns is 3.56% for Bitcoin and 5.10% for Ethereum, confirming that Ethereum is the more volatile of the two, consistent with its smaller market capitalisation and broader exposure to the decentralised finance ecosystem. The minimum observed return for Bitcoin is -51.81%, which corresponds to the COVID-19 sell-off in March 2020, while the corresponding figure for Ethereum is -59.82%.

Table 1 Descriptive Statistics of Daily Log Returns (in Percent), 2016 to 2025.

Asset	N	Mean	Median	Std. Dev.	Min	Max	Skewness	ExcessKurt.	JB Stat.	JB p
Bitcoin	3,652	0.145	0.136	3.563	-51.808	25.092	-0.887	16.335	41,084.0	< 0.001
Ethereum	3,652	0.220	0.063	5.103	-59.825	29.516	-0.340	10.297	16,203.0	< 0.001

Notes: Daily log returns $r_t = 100 \times \ln(P_t / P_{t-1})$ for Bitcoin and Ethereum over 1 January 2016 to 31 December 2025. JB Stat. is the Jarque-Bera test statistic for normality. The null hypothesis of normality is rejected at all conventional significance levels for both series.

Both return series exhibit pronounced negative skewness and large excess kurtosis. The excess kurtosis of 16.34 for Bitcoin and 10.30 for Ethereum indicates that extreme returns occur far more frequently than under the normal distribution, a finding that motivates the use of a Student-t distribution in the parametric models developed in Section 4. The Jarque-Bera test rejects the null of normality at any conventional significance level for both series.

➤ Stylised Facts and Diagnostic Tests

Table 2 reports four standard diagnostic tests that establish the stylised facts of cryptocurrency returns and motivate the modelling approach adopted in the next section. Panel A presents the augmented Dickey-Fuller (ADF) test for stationarity. The null hypothesis of a unit root is decisively rejected for both return series and for both squared-return series, confirming that the series are stationary and suitable for direct modelling without differencing. Panel B reports the Ljung-Box Q statistic for the null of no autocorrelation up to lag 20. While the autocorrelation of the raw returns is weak for Bitcoin and modest for Ethereum, the squared returns

exhibit strong autocorrelation for both assets, with Q(20) statistics of 140.9 for Bitcoin and 357.0 for Ethereum, both significant at the 0.1% level. This pattern is the signature of volatility clustering. Panel C reports Engle's ARCH-LM test for conditional heteroskedasticity. The null of no ARCH effects is rejected at any conventional level for both assets at lags 5 and 10. Together these results justify the use of GARCH-family models for the conditional variance.

Table 2 Stationarity and Serial-Dependence Diagnostics.

Series / Test	Statistic	p-value	Conclusion
Panel A: Augmented Dickey-Fuller test			
BTC log returns	-42.32	< 0.001	Reject unit root
ETH log returns	-11.45	< 0.001	Reject unit root
BTC squared returns	-19.01	< 0.001	Reject unit root
ETH squared returns	-18.05	< 0.001	Reject unit root
Panel B: Ljung-Box Q(20) test			
BTC log returns	26.50	0.150	Marginal
ETH log returns	57.82	< 0.001	Significant
BTC squared returns	140.87	< 0.001	Volatility clustering
ETH squared returns	356.99	< 0.001	Volatility clustering
Panel C: Engle ARCH-LM test (lag 10)			
BTC log returns	89.11	< 0.001	ARCH effects present
ETH log returns	193.72	< 0.001	ARCH effects present

Notes: Panel A reports the augmented Dickey-Fuller test, with the null hypothesis of a unit root. Panel B reports the Ljung-Box Q test at lag 20, with the null of no serial correlation. Panel C reports Engle's ARCH-LM test at lag 10, with the null of no ARCH effects. All tests are based on the full sample from 1 January 2016 to 31 December 2025.

Fig 1 plots the daily closing price, daily log return, and squared daily log return for Bitcoin and Ethereum over the full sample. The price plots reveal the major regimes of the decade: the run-up to the December 2017 peak, the 2018 decline, the 2020 COVID dislocation, the 2021 expansion, the 2022 deleveraging, and the post-spot-ETF environment of 2024 and 2025. The return plots show pronounced volatility clustering, with quiet periods interrupted by bursts of large absolute returns. The squared-return plots make this clustering even more visible and motivate the use of an autoregressive conditional heteroskedasticity model.

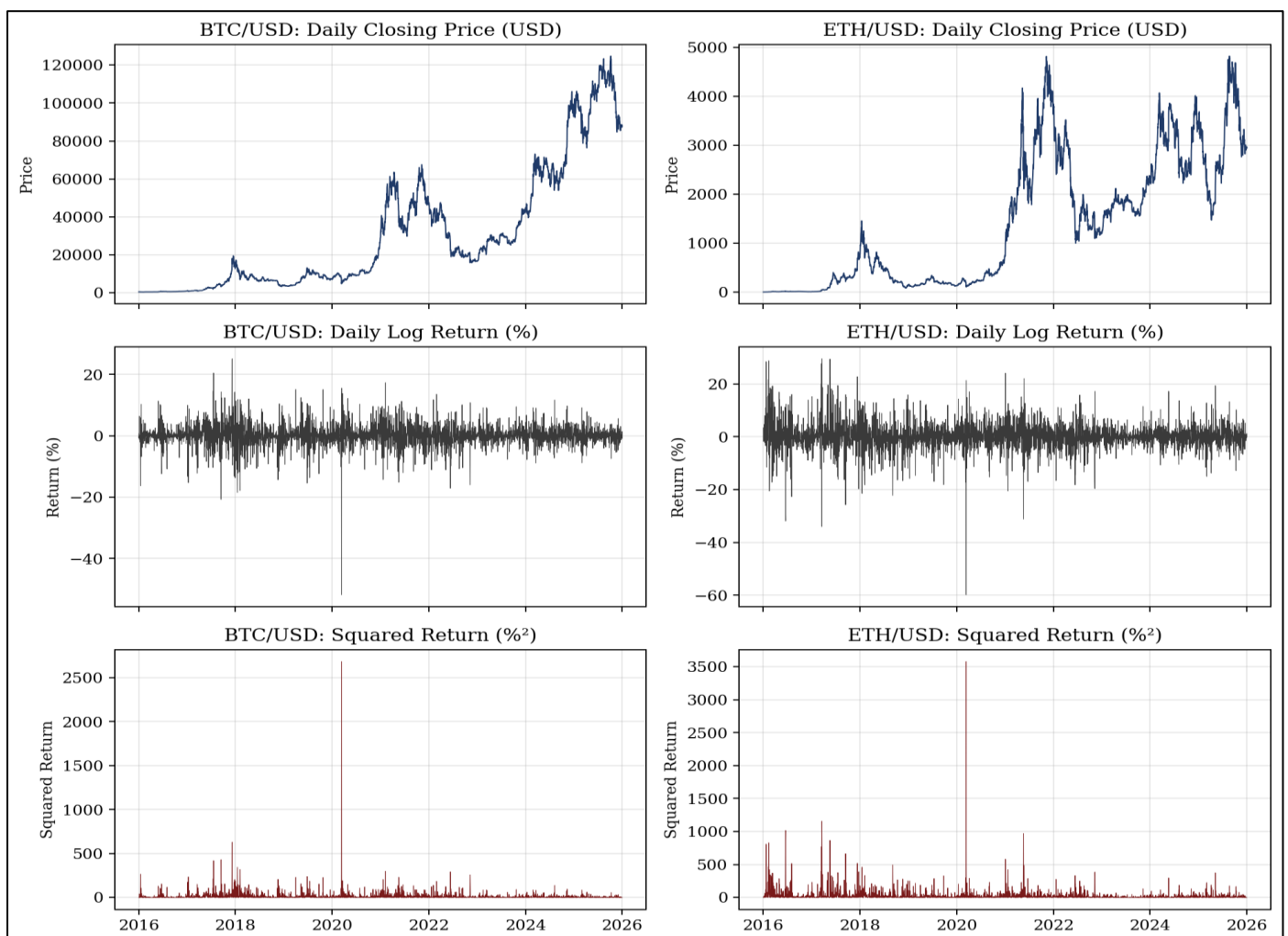


Fig 1 Daily Closing Price, Daily Log Return (in Percent), and Daily Squared Log Return (in Percent Squared) for Bitcoin and Ethereum, 1 January 2016 to 31 December 2025.

Fig 2 displays the autocorrelation functions of the raw and squared returns for both assets up to lag 40. The ACF of raw returns is largely consistent with white noise for Bitcoin and shows mild positive autocorrelation at short lags for Ethereum. The ACF of squared returns, by contrast, decays slowly and remains significantly positive over many lags for both assets, consistent with the long-memory property of cryptocurrency volatility documented by Kaya Soylu et al. (2020) and others.

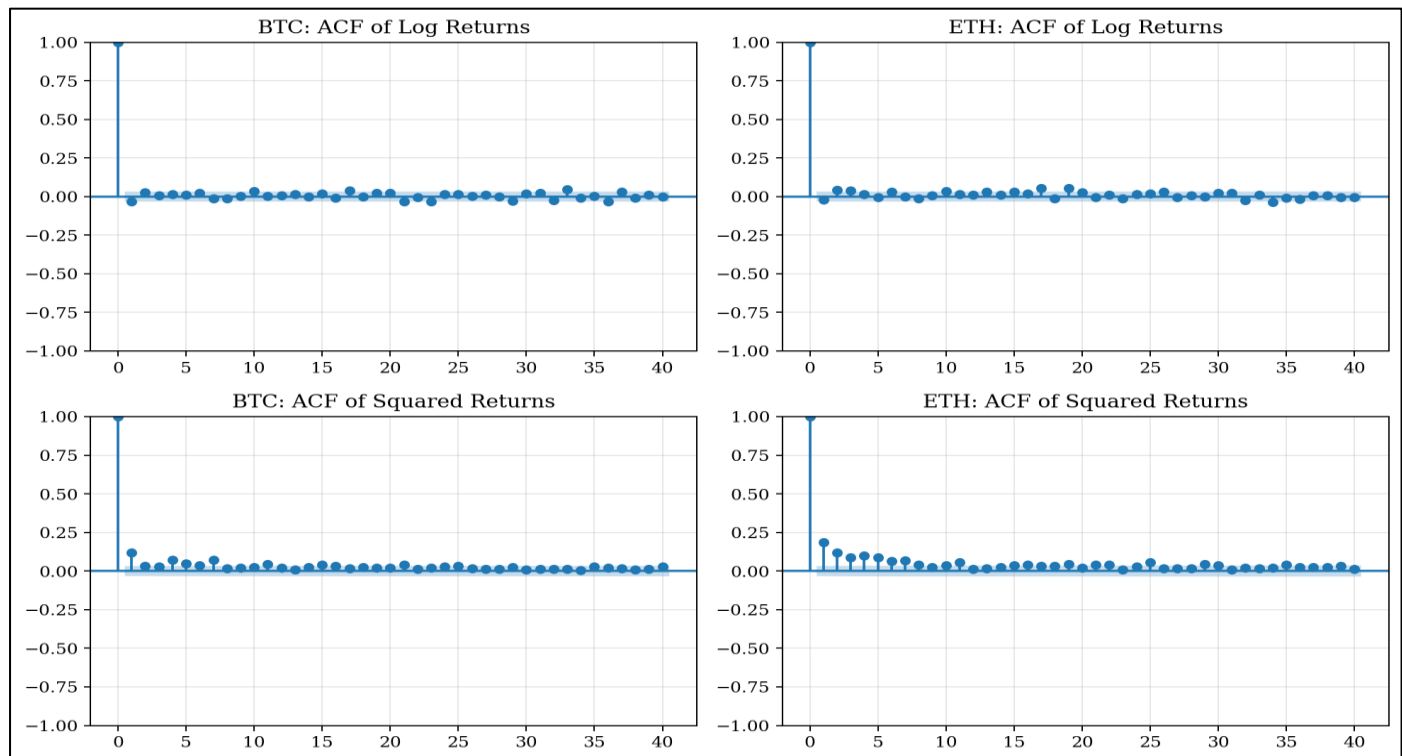


Fig 2 Autocorrelation Functions of Daily Log Returns And Daily Squared Log Returns For Bitcoin And Ethereum. The Shaded Bands Are 95% Confidence Intervals Under The Null Of No Autocorrelation.

IV. METHODOLOGY

➤ The GARCH(1,1) Model

The first parametric benchmark is the symmetric GARCH(1,1) model of Bollerslev (1986). Let r_t denote the daily log return on day t and let ε_t denote the innovation defined by:

$$r_t = \mu + \varepsilon_t, \varepsilon_t = \sigma_t z_t \quad (2)$$

where μ is a constant mean and z_t is an i.i.d. innovation with zero mean and unit variance. The conditional variance σ_t^2 evolves according to:

$$\sigma_t^2 = \omega + \alpha \varepsilon_{t-1}^2 + \beta \sigma_{t-1}^2 \quad (3)$$

with the standard restrictions $\omega > 0$, $\alpha \geq 0$, $\beta \geq 0$, and $\alpha + \beta < 1$ to ensure non-negativity and covariance stationarity. The standardised innovation z_t is assumed to follow a Student-t distribution with ν degrees of freedom to accommodate the heavy tails documented in Section 3. The GARCH model is estimated by maximum likelihood, with robust standard errors computed using the sandwich formula. It serves both as a baseline against which the more elaborate EGARCH and machine learning models are compared, and as a check on whether the asymmetric structure of EGARCH is empirically warranted in this sample.

➤ The EGARCH(1,1) Model

The second parametric model is the Exponential GARCH model of Nelson (1991), which captures asymmetric volatility responses to positive and negative shocks while imposing no positivity restrictions on the parameters. Under the same notation as in equation (2), the EGARCH(1,1) specification for the conditional variance is:

$$\ln(\sigma_t^2) = \omega + \alpha [|z_{t-1}| - E|z_{t-1}|] + \gamma z_{t-1} + \beta \ln(\sigma_{t-1}^2) \quad (4)$$

where the coefficient α governs the response of conditional variance to the magnitude of past shocks, γ captures the asymmetric or leverage effect, and β measures volatility persistence. Modelling the logarithm of the conditional variance ensures that σ_t^2 remains positive without the need for parameter constraints. The model is covariance-stationary whenever $|\beta| < 1$.

As with the GARCH model, the standardised innovation z_t is assumed to follow a Student-t distribution with ν degrees of freedom. The choice of the Student-t distribution is standard in cryptocurrency volatility modelling and is supported by Mohamed et al. (2025) and Sözen (2025). The full set of parameters $\theta = (\mu, \omega, \alpha, \gamma, \beta, \nu)$ is estimated by maximum likelihood. The adequacy of the EGARCH specification is assessed in Section 5 through Ljung-Box and ARCH-LM tests on the standardised residuals; this check is important because the residual-based hybrid (Hybrid-B) inherits any misspecification of the EGARCH model. One-

step-ahead variance forecasts $\hat{\sigma}_{t+1}^2$ are computed recursively from the fitted equation, with the conditional variance updated using the realised innovation ε_t at each new observation.

➤ The LSTM Neural Network

The third model is a Long Short-Term Memory neural network (Hochreiter & Schmidhuber, 1997). The LSTM is a recurrent neural network architecture designed to learn long-range dependencies in sequential data without suffering the vanishing-gradient problem that affects vanilla recurrent networks. At each time step t , the LSTM cell maintains a hidden state h_t and a cell state c_t , which are updated according to a system of gating equations:

$$f_t = \sigma(W_f x_t + U_f h_{t-1} + b_f) \quad (5)$$

$$i_t = \sigma(W_i x_t + U_i h_{t-1} + b_i) \quad (6)$$

$$o_t = \sigma(W_o x_t + U_o h_{t-1} + b_o) \quad (7)$$

$$\tilde{c}_t = \tanh(W_c x_t + U_c h_{t-1} + b_c) \quad (8)$$

$$c_t = f_t \odot c_{t-1} + i_t \odot \tilde{c}_t \quad (9)$$

$h_t = o_t \odot \tanh(c_t)$ (10) where $\sigma(\cdot)$ denotes the sigmoid function, $\odot$ denotes the element-wise product, f_t , i_t , and o_t are the forget, input, and output gates respectively, $\tilde{c}_t$ is the candidate cell state, and $\{W, U, b\}$ are learnable weight matrices and bias vectors. The forget gate controls how much of the previous cell state is retained, the input gate determines how much new information is incorporated, and the output gate regulates the contribution of the cell state to the hidden state. This gating mechanism allows the LSTM to maintain useful information over many time steps and to discard irrelevant signals (Staudemeyer & Morris, 2019).

In the present application, the input vector x_t at each time step contains the lagged daily log return and the lagged squared log return, both standardised using the training-sample mean and standard deviation. The lookback window is set to twenty-two trading days, corresponding approximately to one calendar month and consistent with the lookback used by Pratas et al. (2023) and Phan Huy et al. (2025). The architecture consists of two stacked LSTM layers with thirty-two hidden units each, followed by dropout layers with rate 0.2 to mitigate overfitting, a dense layer with sixteen units and ReLU activation, and a final scalar output passed through a softplus activation to ensure that the predicted variance is non-negative. The target variable is the next-day squared log return, transformed by $\log(1 + y)$ to stabilise variance, and then standardised. The model is trained with the Adam optimiser, mean squared error loss, early stopping on a validation set with patience of fifteen epochs, and a maximum of one hundred epochs.

Because the parameters of a recurrent neural network are sensitive to the random initialisation of weights and to the order in which training samples are presented (Reimers & Gurevych, 2017), the LSTM and both hybrids are estimated under ten different random seeds, namely $\{7, 11, 42, 99, 123,$

$271, 314, 555, 888, 2024\}$. For each seed, the entire training and forecasting procedure is repeated end-to-end. The headline forecast for each model and each test date is the median across the ten seeds, with the seed-to-seed dispersion of the evaluation metrics reported separately as a measure of model stability.

➤ Hybrid-A: EGARCH Variance as LSTM Feature

The first hybrid specification, denoted Hybrid-A, augments the LSTM input with the one-step-ahead conditional variance forecast produced by the EGARCH model. At each time step, the input vector becomes:

$$x_t = (r_{t-1}, r_{t-1}^2, \hat{\sigma}_t^2)^T \quad (11)$$

where $\hat{\sigma}_t^2$ is the EGARCH conditional variance forecast for day t made on the basis of information up to day $t - 1$. This construction provides the LSTM with a parametric volatility signal alongside the raw return information and follows the design adopted by Kristjanpoller and Minutolo (2018), Seo and Kim (2020), and Zahid et al. (2022). The remaining architectural details, including the lookback window, hidden-unit count, dropout, and training procedure, are identical to those of the pure LSTM described in Section 4.3 to ensure a fair comparison.

➤ Hybrid-B: LSTM on EGARCH Standardised Residuals

The second hybrid specification, denoted Hybrid-B, takes a different route. Rather than feeding the EGARCH variance forecast as an input feature, it filters the return series through the EGARCH model to obtain standardised residuals:

$$\tilde{z}_t = (r_t - \mu) / \hat{\sigma}_t \quad (12)$$

and trains the LSTM on $(\tilde{z}_{t-1}, \tilde{z}_{t-1}^2)$. The intuition is that the EGARCH model captures the autoregressive component of conditional variance, while the LSTM learns to detect the residual nonlinear structure in the standardised innovations, in the spirit of Araya et al. (2024) and Zhao et al. (2025). When the LSTM operates on a series that has already been filtered for the dominant volatility dynamics, any improvement over the pure LSTM benchmark must come from genuine residual nonlinearity that the parametric model failed to capture.

➤ Forecasting Procedure

All models are evaluated under an expanding-window rolling forecast scheme. The training window starts on 1 January 2016 and the test window runs from 1 January 2023 to 31 December 2025, providing 1,096 daily out-of-sample forecasts per model and per asset. For the GARCH and EGARCH models, the parameters are re-estimated every thirty days using the full history available up to that point, balancing computational cost against the risk of parameter drift; one-step-ahead variance forecasts between refits are produced by recursive updating with realised innovations. For the LSTM and the two hybrid models, the network is trained once on data up to the start of the test window and used to produce sequential one-step-ahead forecasts thereafter, with no re-training during the test period; this design choice mirrors the standard practice in the LSTM

volatility literature (Pratas et al., 2023; Dudek et al., 2024). A fifteen-percent slice of the pre-test data, taken from the most recent end of the training sample, serves as the validation set for early stopping. The fifteen-percent figure is consistent with conventions in the time-series deep learning literature and was selected because the validation set is used only to trigger early stopping rather than for model selection or hyperparameter tuning, so a relatively small slice is sufficient. Sensitivity of the headline results to this choice is examined in Section 5.7 below, where the pure LSTM and Hybrid-B are shown to vary by less than 0.71 percent in MSE across the 75/25, 80/20, and 85/15 configurations.

➤ Evaluation Framework

Forecast accuracy is assessed under five complementary metrics. The first three are the mean squared error, the mean absolute error, and the root mean squared error, each computed by comparing the variance forecast to the realised squared return. The fourth is the QLIKE loss of Patton (2011), which is robust to noise in the squared-return proxy and penalises underestimation of variance more heavily than overestimation:

$$\text{QLIKE}(\sigma_t^2, r_t^2) = \ln(\sigma_t^2) + r_t^2 / \sigma_t^2 \quad (13)$$

The fifth is the Mincer-Zarnowitz coefficient of determination obtained from regressing the realised squared return on the variance forecast and a constant. While Mincer-Zarnowitz R^2 values are mechanically small when the dependent variable is the noisy squared-return proxy, they remain useful for cross-model comparison.

Pairwise predictive accuracy is assessed using the Diebold and Mariano (1995) test with the small-sample correction of Harvey, Leybourne, and Newbold (1997). The test statistic is computed under each loss function and compared to the t-distribution with $n - 1$ degrees of freedom, where n is the number of out-of-sample observations. To control for the multiple comparison problem inherent in the simultaneous evaluation of six models, the Model Confidence Set procedure of Hansen, Lunde, and Nason (2011) is applied at the 90% confidence level. The procedure returns the set of

models that are not statistically distinguishable from the best at the chosen significance level. Critical values are obtained by stationary bootstrap (Politis & Romano, 1994) with $B = 1,000$ replications and an expected block length of 10 days. The Model Confidence Set is computed separately under the MSE and QLIKE losses to highlight the sensitivity of the rankings to the choice of loss function.

V. EMPIRICAL RESULTS

➤ GARCH and EGARCH Parameter Estimates

Table 3 reports the maximum-likelihood parameter estimates for the GARCH(1,1)-t and EGARCH(1,1)-t models fitted on the full sample. The estimates of α (the ARCH effect) and β (the persistence parameter) are highly significant for both models and both assets, confirming the presence of strong conditional heteroskedasticity and volatility clustering. The estimated persistence parameter β is 0.978 for the GARCH model on Bitcoin and 0.991 for the EGARCH model, with corresponding values of 0.945 and 0.966 for Ethereum, indicating near-integrated volatility dynamics consistent with the long-memory findings of Kaya Soyly et al. (2020). The Student-t degrees-of-freedom parameter ν is 2.65 for the EGARCH model on Bitcoin and 3.05 for Ethereum, again confirming heavy tails and supporting the choice of the Student-t distribution.

An interesting feature of the EGARCH estimates is that the asymmetric parameter γ is small and only marginally significant for Bitcoin (estimate 0.024, $p = 0.083$) and statistically insignificant for Ethereum (estimate 0.002, $p = 0.922$). The leverage effect that motivates the EGARCH specification in equity markets therefore appears muted in this cryptocurrency sample. EGARCH still attains a slightly lower AIC than GARCH for both assets (BTC: 17,957 vs 18,005; ETH: 20,769 vs 20,795), consistent with the marginal relevance of γ at the in-sample level, but the parsimony of GARCH makes it a serious competitor for out-of-sample forecasting. This finding mirrors the asymmetry results of Cheikh et al. (2020) and motivates the explicit comparison of the two parametric specifications in subsequent sections.

Table 3 Full-Sample Maximum-Likelihood Estimates Of GARCH(1,1)-T And EGARCH(1,1)-T Models.

Parameter	BTC	(s.e.)	ETH	(s.e.)	BTC	(s.e.)	ETH	(s.e.)
Panel A: Mean equation								
μ	0.131	(0.032)	0.091	(0.050)	0.125	(0.030)	0.077	(0.051)
Panel B: Variance equation								
ω	0.150	(0.080)	1.198	(0.465)	0.079	(0.021)	0.158	(0.051)
α	0.096	(0.013)	0.185	(0.036)	0.237	(0.031)	0.288	(0.051)
γ (asymmetry)	—	—	—	—	0.024	(0.014)	0.002	(0.017)
β	0.904	(0.018)	0.815	(0.040)	0.991	(0.005)	0.966	(0.013)
ν	3.174	(0.130)	3.152	(0.164)	2.653	(0.142)	3.051	(0.179)
Panel C: Fit statistics								

Log-likelihood	-8997.4		-0392.6		-8972.5		-10378.5	
AIC	18004.8		20795.2		17956.9		20769.0	
BIC	18035.8		20826.2		17994.1		20806.2	
N	3,652		3,652		3,652		3,652	

Notes: Maximum likelihood estimates with robust standard errors in parentheses. Columns 2-5 report GARCH(1,1)-t estimates and columns 6-9 report EGARCH(1,1)-t estimates. The full sample runs from 1 January 2016 to 31 December 2025. The asymmetry parameter γ applies only to the EGARCH specification.

➤ Adequacy of the Parametric Models

Before turning to forecast comparison, Table 4 reports diagnostic tests on the standardised residuals from the fitted GARCH and EGARCH models. The Ljung-Box statistic computed on the squared standardised residuals tests for any remaining autoregressive conditional heteroskedasticity, and Engle's ARCH-LM statistic provides a complementary check. For both models and both assets the null hypothesis of no remaining ARCH effects cannot be rejected, with p-values comfortably above 0.40 in every case. This indicates that both parametric specifications adequately capture the volatility

dynamics of the return series and provide a well-specified foundation for the residual-based hybrid (Hybrid-B). The Ljung-Box test on the standardised residuals themselves (rather than their squares) detects modest residual autocorrelation in the mean equation, particularly for Ethereum ($Q(10) = 28.2$, $p = 0.002$ for EGARCH); a richer mean specification might absorb this autocorrelation but is not pursued here, since the focus is on conditional variance forecasting.

Table 4 Diagnostic Tests on Standardised Residuals from GARCH(1,1)-T And EGARCH(1,1)-T Models.

Series and Test	Statistic	p-value	Significance	Conclusion
Panel A: BTC				
GARCH - Ljung-Box $Q(10)$ on z^2	6.09	0.807	0.05	Cannot reject
GARCH - Ljung-Box $Q(20)$ on z^2	11.11	0.943	0.05	Cannot reject
GARCH - ARCH-LM(5)	4.89	0.429	0.05	Cannot reject
GARCH - ARCH-LM(10)	6.11	0.806	0.05	Cannot reject
EGARCH - Ljung-Box $Q(10)$ on z^2	7.77	0.651	0.05	Cannot reject
EGARCH - Ljung-Box $Q(20)$ on z^2	11.02	0.946	0.05	Cannot reject
EGARCH - ARCH-LM(5)	7.16	0.209	0.05	Cannot reject
EGARCH - ARCH-LM(10)	7.67	0.661	0.05	Cannot reject
Panel B: ETH				
GARCH - Ljung-Box $Q(10)$ on z^2	9.83	0.456	0.05	Cannot reject
GARCH - Ljung-Box $Q(20)$ on z^2	18.93	0.526	0.05	Cannot reject
GARCH - ARCH-LM(5)	3.65	0.600	0.05	Cannot reject
GARCH - ARCH-LM(10)	10.13	0.429	0.05	Cannot reject
EGARCH - Ljung-Box $Q(10)$ on z^2	7.70	0.658	0.05	Cannot reject
EGARCH - Ljung-Box $Q(20)$ on z^2	16.15	0.707	0.05	Cannot reject
EGARCH - ARCH-LM(5)	2.50	0.776	0.05	Cannot reject
EGARCH - ARCH-LM(10)	7.71	0.657	0.05	Cannot reject

Notes: Diagnostic tests on standardised residuals $z_t = \varepsilon_t/\sigma_t$ from full-sample GARCH(1,1)-t and EGARCH(1,1)-t fits. Ljung-Box tests are computed on the squared standardised residuals z^2 and test the null of no remaining ARCH effects. ARCH-LM(k) is Engle's Lagrange multiplier test at lag k. The conclusion column refers to the null hypothesis of no remaining ARCH effects at the 5% level.

➤ Out-of-Sample Forecast Accuracy

Tables 5 and 6 report the out-of-sample forecast accuracy of the seven models for Bitcoin and Ethereum, respectively. The seven models are the symmetric GARCH(1,1)-t, the asymmetric EGARCH(1,1)-t, the pure LSTM, the two hybrid LSTMs, a constant unconditional-variance benchmark equal to the pre-test sample mean of squared returns, and the naive random-walk benchmark that uses the previous day's squared return as the next day's variance forecast. The LSTM, Hybrid-A, and Hybrid-B forecasts are the median across the ten random seeds described in Section 4.3. Inclusion of the constant benchmark addresses a classic concern in the volatility forecasting literature: that a model achieving low MSE on a noisy proxy may simply be predicting the unconditional mean rather than

capturing genuine volatility dynamics. If the constant benchmark beats the structured models, the apparent victory of any of those models is suspect.

For Bitcoin, the LSTM achieves the lowest MSE at 157.78, followed almost identically by Hybrid-B at 157.87. The two predictions differ by only 0.09, which is within the seed-to-seed noise of either model and well below the gap to any other model. In what follows we treat Hybrid-B as effectively collapsing to the pure LSTM rather than as a distinct alternative. Hybrid-A is noticeably higher at 162.99. The two parametric models trail further behind, with GARCH at 169.36 and EGARCH at 262.51. The constant benchmark sits at 251.35, which is comfortably above the LSTM and GARCH but actually below EGARCH; in other words,

EGARCH produces forecasts no better than predicting the unconditional mean every day. The naive random-walk benchmark is the worst performer at 267.38. The relative ordering on MAE is similar, with the LSTM and Hybrid-B essentially tied at 6.12 and the other models trailing. Under the QLIKE loss the picture is more compressed: the LSTM, Hybrid-B, and Hybrid-A all cluster around 2.83, with GARCH (2.892) within striking distance and EGARCH (2.986) and the constant (3.124) further behind. The naive benchmark is dramatically penalised under QLIKE (575.86) because of the heavy penalty on near-zero variance forecasts.

For Ethereum, the LSTM and Hybrid-B again produce essentially identical forecasts, with MSE of 671.63 and 671.69 respectively. EGARCH and GARCH are essentially tied at 747.58 and 730.10, and Hybrid-A is the worst-performing of the structured forecasts at 815.19. The constant benchmark is the next-worst at 1,152.30, and the naive benchmark is the weakest at 1,190.19. Importantly, every

structured model on Ethereum beats the constant benchmark by a wide margin (a 42% reduction in MSE for the LSTM), which provides direct evidence that the LSTM is capturing genuine volatility dynamics and not merely predicting the unconditional mean. The Mincer-Zarnowitz R^2 values reported in the rightmost column are very low across all models, ranging from 0.0001 to 0.018 for both assets. These small magnitudes are typical of daily volatility regressions on the squared-return proxy and reflect the high noise level of the proxy rather than weak forecast quality; comparable values are reported by Patton (2011) and Bergsli et al. (2022). Under QLIKE for Ethereum, the LSTM, Hybrid-B, EGARCH, and GARCH all cluster within a narrow band around 3.41 to 3.44, with Hybrid-A (3.564) noticeably worse and the constant (3.812) and naive (1,693.19) benchmarks much weaker. These patterns suggest that under MSE the LSTM-based models clearly dominate, while under QLIKE the parametric and deep learning approaches are far closer.

Table 5 Out-of-Sample Forecast Accuracy: Bitcoin (Test Period 1 January 2023 To 31 December 2025).

Model	MSE	MAE	RMSE	QLIKE	R^2 (MZ)
GARCH(1,1)-t	169.36	7.660	13.014	2.892	0.0002
EGARCH(1,1)-t	262.51	11.499	16.202	2.986	0.0001
LSTM (median)	157.78	6.117	12.561	2.834	0.0106
Hybrid-A (median)	162.99	7.014	12.767	2.833	0.0016
Hybrid-B (median)	157.87	6.115	12.565	2.837	0.0016
Constant (uncond.)	251.35	13.713	15.854	3.124	0.0000
Naive (RW)	267.38	8.076	16.352	575.855	0.0185

Notes: Out-of-sample forecast accuracy for Bitcoin against the squared-return proxy. LSTM, Hybrid-A, and Hybrid-B are reported as the median across ten random seeds {7, 11, 42, 99, 123, 271, 314, 555, 888, 2024}. The Constant benchmark predicts the pre-test sample mean of squared returns (15.66) every day. The naive benchmark uses the previous day's squared return. MSE and RMSE are in (percent squared) squared; MAE is in percent squared; QLIKE is unitless; R^2 (MZ) is the Mincer-Zarnowitz coefficient of determination. Low R^2 values are typical of daily volatility regressions on the squared-return proxy. $n = 1,096$.

Table 6 Out-of-Sample Forecast Accuracy: Ethereum (Test Period 1 January 2023 To 31 December 2025).

Model	MSE	MAE	RMSE	QLIKE	R^2 (MZ)
GARCH(1,1)-t	730.10	16.072	27.020	3.436	0.0001
EGARCH(1,1)-t	747.58	16.735	27.342	3.422	0.0027
LSTM (median)	671.63	11.363	25.916	3.413	0.0004
Hybrid-A (median)	815.19	20.704	28.552	3.564	0.0063
Hybrid-B (median)	671.69	11.360	25.917	3.413	0.0001
Constant (uncond.)	1152.28	28.711	33.945	3.812	-0.0000
Naive (RW)	1190.19	15.061	34.499	1693.194	0.0108

Notes: Out-of-sample forecast accuracy for Ethereum against the squared-return proxy, computed analogously to Table 5. The Constant benchmark for Ethereum predicts the pre-test sample mean of squared returns (32.72) every day. The high QLIKE value for the naive benchmark reflects the heavy penalty that QLIKE applies when the variance forecast is close to zero, which occurs whenever the previous day's return is close to zero.

Fig 3 visualises the variance forecasts of all five model-based methods alongside the realised squared returns over the test period, on a logarithmic vertical axis. Two patterns are visible. First, the LSTM and Hybrid-B forecasts are nearly identical and form a roughly horizontal line around the unconditional mean variance, with only mild variation. Second, the GARCH and EGARCH forecasts respond visibly to volatility regimes, stepping up during episodes of elevated

realised variance and stepping down during quieter periods. Hybrid-A largely tracks the EGARCH path, reflecting the strong influence of the EGARCH variance feature on the LSTM input. The visual contrast between the responsive parametric forecasts and the near-constant LSTM forecasts is consistent with the observation that the LSTM minimises MSE by predicting close to the unconditional mean of the noisy squared-return target.

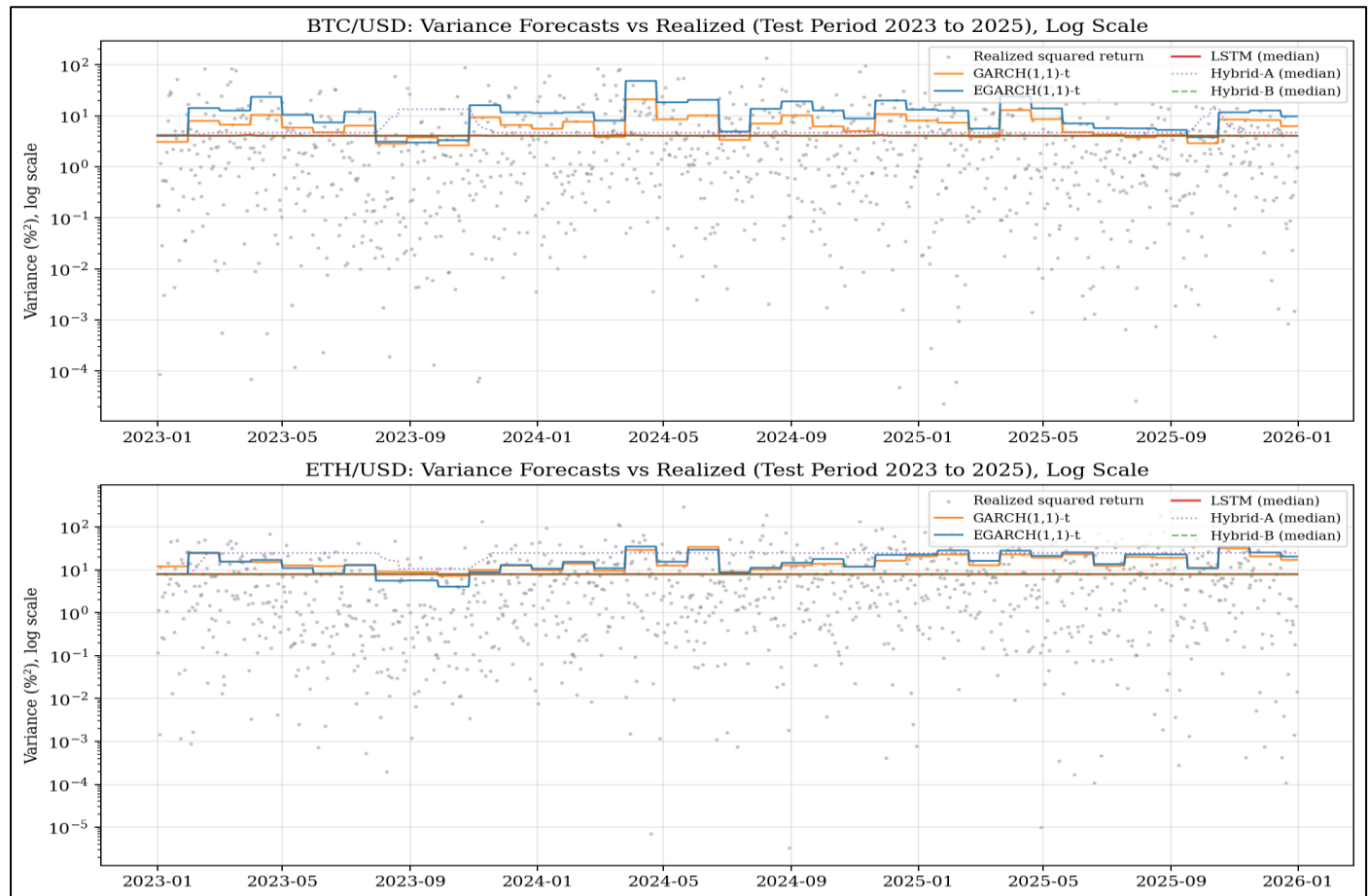


Fig 3 Variance Forecasts and Realised Squared Returns over The Test Period (1 January 2023 To 31 December 2025) on A Logarithmic Scale. The LSTM, Hybrid-A, And Hybrid-B Forecasts Are The Median Across Ten Random Seeds.

Fig 4 plots the cumulative squared-error loss over time for each model. The cumulative loss curves of the LSTM and Hybrid-B coincide almost exactly throughout the test period, while the GARCH and Hybrid-A curves are slightly above and the EGARCH curve is well above. All curves accelerate during the high-volatility year of 2024 and flatten during the calmer stretches, showing that the bulk of the cumulative MSE difference between models is concentrated in a small number of high-impact days.

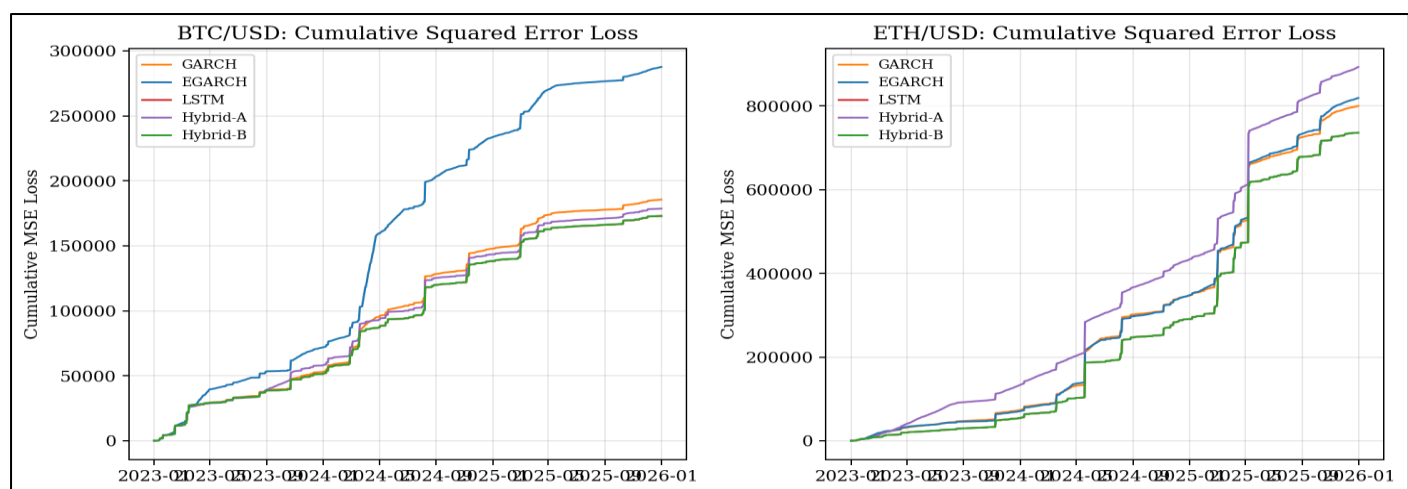


Fig 4 Cumulative Squared-Error Loss for Each Model Over The Test Period 2023 To 2025.

➤ Diebold-Mariano Tests and Model Confidence Set

Tables 7 and 8 report Diebold-Mariano test results for the principal pairwise comparisons under the MSE and QLIKE losses, respectively. Under MSE, the LSTM and Hybrid-B are statistically indistinguishable for both Bitcoin

(DM-HLN $p = 0.42$) and Ethereum ($p = 0.91$), confirming the visual impression from Figure 3. Each of the LSTM-based models significantly outperforms the parametric models for both assets at the 1% level. The two parametric models are themselves close to one another: GARCH significantly

outperforms EGARCH for Bitcoin ($p < 0.001$) but the two are statistically indistinguishable for Ethereum ($p = 0.43$). Under QLIKE, the picture changes substantially. For Bitcoin, GARCH is statistically indistinguishable from the LSTM ($p = 0.41$) and from Hybrid-B ($p = 0.39$), although both LSTM-

based models reject equality with EGARCH ($p < 0.01$). For Ethereum, GARCH, EGARCH, the LSTM, and Hybrid-B form a tightly clustered group in which only Hybrid-A and the naive benchmark stand apart.

Table 7 Diebold-Mariano Test P-Values under MSE Loss (HLN-Corrected).

Model	GARCH	EGARCH	LSTM	Hybrid-A	Hybrid-B
Panel A: BTC					
GARCH	—	< 0.001	< 0.001	0.073	< 0.001
EGARCH	< 0.001	—	< 0.001	< 0.001	< 0.001
LSTM	< 0.001	< 0.001	—	0.017	0.002
Hybrid-A	0.073	< 0.001	0.017	—	0.019
Hybrid-B	< 0.001	< 0.001	0.002	0.019	—
Panel B: ETH					
GARCH	—	< 0.001	< 0.001	< 0.001	< 0.001
EGARCH	< 0.001	—	< 0.001	< 0.001	< 0.001
LSTM	< 0.001	< 0.001	—	< 0.001	0.003
Hybrid-A	< 0.001	< 0.001	< 0.001	—	< 0.001
Hybrid-B	< 0.001	< 0.001	0.003	< 0.001	—

Notes: Reported entries are p-values from the Diebold-Mariano test of equal predictive accuracy under the MSE loss, with the Harvey-Leybourne-Newbold small-sample correction. Each cell (i, j) tests the null that model i and model j have equal expected loss; rejection at the 5% level implies that the row model is significantly different from the column model.

Table 8 Diebold-Mariano test p-values under QLIKE loss (HLN-corrected).

Model	GARCH	EGARCH	LSTM	Hybrid-A	Hybrid-B
Panel A: BTC					
GARCH	—	0.004	0.103	0.156	0.121
EGARCH	0.004	—	0.005	0.003	0.007
LSTM	0.103	0.005	—	0.963	0.002
Hybrid-A	0.156	0.003	0.963	—	0.874
Hybrid-B	0.121	0.007	0.002	0.874	—
Panel B: ETH					
GARCH	—	0.155	0.620	< 0.001	0.628
EGARCH	0.155	—	0.863	< 0.001	0.869
LSTM	0.620	0.863	—	0.023	0.003
Hybrid-A	< 0.001	< 0.001	0.023	—	0.023
Hybrid-B	0.628	0.869	0.003	0.023	—

Notes: Reported entries are p-values from the Diebold-Mariano test of equal predictive accuracy under the QLIKE loss, with the Harvey-Leybourne-Newbold small-sample correction. With ten pairwise comparisons per panel, the family-wise error rate calls for an adjustment in principle. Applying the Holm-Bonferroni procedure to all ten pairs leaves all p-values that are reported here as below 0.001 still significant at the 5% family-wise level, while borderline values (between 0.05 and 0.20) are not. The substantive conclusions in the text rely on the Model Confidence Set, which controls family-wise error by construction.

Table 9 reports the Model Confidence Set results at the 90% confidence level under both losses, with all seven models in the candidate set. Under MSE, the procedure retains a single model for Ethereum, namely the LSTM, with the next-best Hybrid-B eliminated at $p = 0.017$. For Bitcoin, the MCS retains the three LSTM-based models, with the GARCH baseline eliminated at $p = 0.049$, very close to the 90% threshold. Under QLIKE, the parametric models retain a place in the confidence set. For Ethereum, GARCH, EGARCH, the LSTM, and Hybrid-B all survive, with only Hybrid-A, the constant benchmark, and the naive random walk eliminated. For Bitcoin, GARCH survives alongside the three LSTM-based models, with EGARCH eliminated at $p = 0.062$ (close to the boundary). The constant benchmark and the naive random walk are eliminated in every panel. These results crystallise the principal message of the paper: the

choice of loss function fundamentally alters the conclusions one would draw about the relative merits of econometric and deep learning approaches in cryptocurrency volatility forecasting.

Table 9 Model Confidence Set at 90% confidence level under MSE and QLIKE losses.

Asset	Loss	Models in MCS (90%)	Elimination order (p-value)
BTC	MSE	LSTM (median), Hybrid-A (median), Hybrid-B (median)	Constant (uncond.) (p=0.001); Naive (RW) (p=0.003); EGARCH(1,1)-t (p=0.005); GARCH(1,1)-t (p=0.049)
BTC	QLIKE	GARCH(1,1)-t, LSTM (median), Hybrid-A (median), Hybrid-B (median)	Naive (RW) (p=0.037); Constant (uncond.) (p=0.007); EGARCH(1,1)-t (p=0.062)
ETH	MSE	LSTM (median)	Constant (uncond.) (p=0.000); Naive (RW) (p=0.002); Hybrid-A (median) (p=0.000); EGARCH(1,1)-t (p=0.007); GARCH(1,1)-t (p=0.015); Hybrid-B (median) (p=0.017)
ETH	QLIKE	GARCH(1,1)-t, EGARCH(1,1)-t, LSTM (median), Hybrid-B (median)	Naive (RW) (p=0.080); Constant (uncond.) (p=0.000); Hybrid-A (median) (p=0.015)

Notes: Model Confidence Set computed via stationary bootstrap with $B = 1,000$ replications and an expected block length of 10 days. The MCS at confidence level $(1 - \alpha)$ contains the set of models that cannot be ruled out as the best at the chosen significance level $\alpha = 0.10$. Models are eliminated sequentially based on the maximum-t statistic.

➤ Robustness: Seed Sensitivity and Sub-Period Stability

A core methodological concern in the deep learning literature is that LSTM forecasts can vary substantially with the random seed used to initialise weights. Table 10 reports the distribution of MSE and QLIKE values across the ten random seeds for the three LSTM-based models. The pure LSTM and Hybrid-B are highly stable: for Bitcoin, MSE varies from 157.49 to 158.09 across seeds (a standard deviation of 0.16), and for Ethereum from 670.81 to 671.88 (standard deviation 0.32). Hybrid-A, by contrast, exhibits substantially higher seed-to-seed variation: MSE for Ethereum ranges from 664.29 to 1,417.55 across seeds, with

a standard deviation of 250.41. Similar patterns hold for MAE and QLIKE. This instability is itself diagnostic. Hybrid-A's reliance on the EGARCH variance feature appears to introduce a high-dimensional, nonlinear interaction with the raw return inputs that is harder to learn consistently across initialisations. The headline figures for Hybrid-A reported in Tables 5 and 6 are based on the median across seeds, but the dispersion shown in Table 10 implies that any single training run of Hybrid-A delivers an unreliable forecast. A practitioner adopting this hybrid would face material implementation risk.

Table 10 Seed-to-Seed Distribution of Forecast Metrics Across Ten Random Seeds.

Model and Metric	Mean	Median	Std. Dev.	Min	Max
Panel A: BTC					
LSTM - MSE	157.749	157.762	0.160	157.491	158.089
LSTM - MAE	6.122	6.123	0.010	6.107	6.135
LSTM - QLIKE	2.833	2.834	0.003	2.828	2.839
Hybrid-A - MSE	164.600	163.649	3.766	160.343	172.343
Hybrid-A - MAE	7.085	6.971	0.383	6.656	7.996
Hybrid-A - QLIKE	2.842	2.844	0.016	2.817	2.875
Hybrid-B - MSE	157.906	157.871	0.202	157.714	158.430
Hybrid-B - MAE	6.118	6.117	0.011	6.097	6.132
Hybrid-B - QLIKE	2.838	2.837	0.005	2.832	2.848
Panel B: ETH					
LSTM - MSE	671.511	671.637	0.324	670.809	671.877
LSTM - MAE	11.373	11.362	0.026	11.346	11.431
LSTM - QLIKE	3.412	3.413	0.003	3.406	3.415
Hybrid-A - MSE	954.792	816.699	250.414	664.289	1417.546
Hybrid-A - MAE	22.820	20.736	6.214	11.909	32.803
Hybrid-A - QLIKE	3.620	3.568	0.161	3.362	3.885
Hybrid-B - MSE	671.557	671.690	0.324	670.863	671.889
Hybrid-B - MAE	11.371	11.360	0.026	11.345	11.427
Hybrid-B - QLIKE	3.412	3.413	0.003	3.407	3.415

Notes: Distribution of forecast accuracy metrics across ten random seeds for the three LSTM-based models. Each row reports descriptive statistics of the metric across the ten end-to-end training and forecasting runs, with seeds {7, 11, 42, 99, 123, 271, 314, 555, 888, 2024}.

Fig 5 visualises the seed-distribution information for MSE. The boxes for the LSTM and Hybrid-B are visually narrow and overlap one another, while the box for Hybrid-A is much wider, with the upper whisker for Ethereum extending more than twice as far above the median as below it. The figure underscores that single-seed reporting in the volatility-LSTM literature can mask important variability for some hybrid architectures.

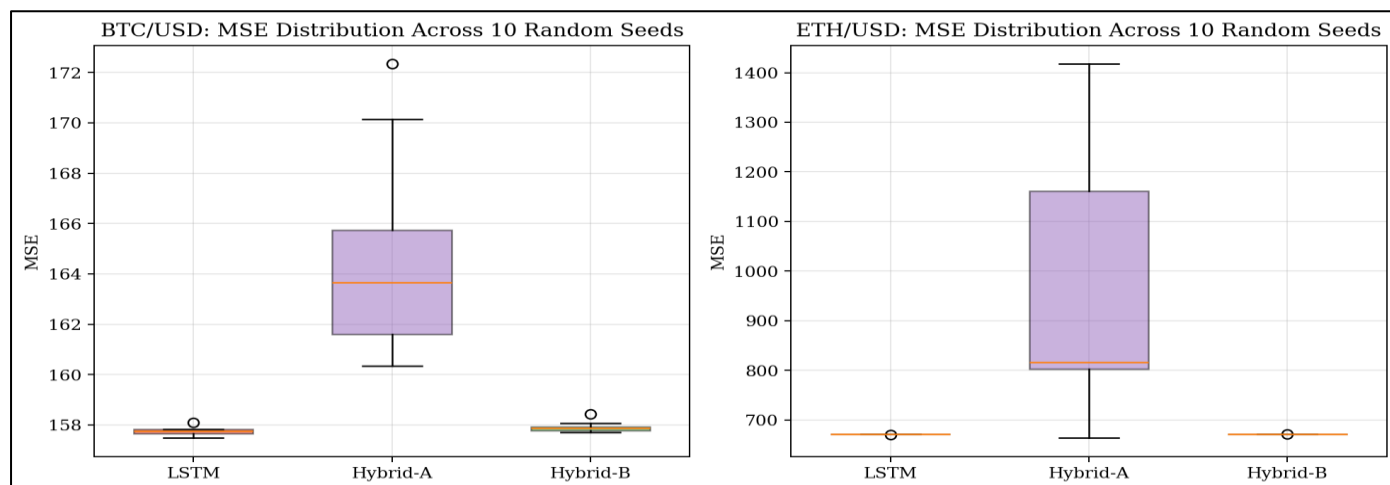


Fig 5 Distribution of Out-of-Sample MSE Across Ten Random Seeds for The Three LSTM-Based Models. The Boxes Show The Inter-Quartile Range, The Central Line Is The Median, And The Whiskers Extend to The Extreme Values; Outliers Are Plotted As Circles.

A second robustness check examines whether the model rankings are stable across sub-periods. Table 11 reports the MSE for each model in each calendar year of the test window. The patterns are consistent with the headline results across all three years for both assets. The LSTM and Hybrid-B produce essentially identical MSE in every year and are the best two models in every year for both assets. GARCH consistently outperforms EGARCH and tracks Hybrid-A reasonably

closely. Volatility levels themselves vary across years, with 2024 being a particularly volatile year for both assets (consistent with the launch of the US spot ETFs in early 2024) and 2025 representing a return to slightly more elevated volatility for Ethereum than for Bitcoin. The relative ordering of the models is preserved across all three regimes, providing reassurance that the headline results are not driven by a single sub-period.

Table 11 Sub-period analysis of out-of-sample MSE by calendar year.

Model	2023	2024	2025
Panel A: BTC			
GARCH	144.97	259.30	103.57
EGARCH	196.94	441.85	148.23
LSTM	140.73	237.24	95.15
Hybrid-A	158.69	233.53	96.56
Hybrid-B	140.75	237.47	95.16
Panel B: ETH			
GARCH	200.31	748.54	1241.41
EGARCH	193.07	757.51	1292.14
LSTM	148.16	646.53	1220.27
Hybrid-A	364.58	820.05	1260.92
Hybrid-B	148.15	646.57	1220.41

Notes: Out-Of-Sample MSE by Calendar Year of the Test Window. Each Year Contains Approximately 365 Daily Forecasts. The Model Ranking is Stable Across All Three Years for Both Assets.

➤ Robustness Across Alternative Volatility Proxies

A final robustness check addresses the noise in the squared-return proxy. Liu et al. (2015) show that the squared daily return is an unbiased but extremely noisy estimator of latent variance, and that range-based and realised-variance estimators dominate it on signal-to-noise grounds. The OHLC data used in this study permit the construction of two such alternative estimators: the Parkinson (1980) estimator and the Garman and Klass (1980) estimator, both described in Section 3.1. Table 12 reports the out-of-sample MSE of all seven models against each of the three proxies for both assets. Two patterns are notable. First, the absolute magnitude of MSE drops sharply when the cleaner range-based proxies are used, as expected: for Bitcoin, the LSTM's MSE falls from

157.78 against the squared-return proxy to 69.26 against Parkinson and 65.10 against Garman-Klass; for Ethereum, the corresponding figures are 671.63, 245.66, and 220.11. The reduction reflects the lower noise variance of the range-based proxies rather than any change in forecast quality. Second, and more important, the relative ranking of the seven models is preserved across all three proxies. The LSTM and Hybrid-B are tied for the best MSE under every proxy and for both assets; Hybrid-A trails them under every proxy; GARCH is between the LSTM and EGARCH under every proxy; and EGARCH, the constant benchmark, and the naive random walk occupy the bottom tier. The Bitcoin Parkinson and Garman-Klass results sharpen one observation made earlier: the EGARCH MSE of 171.74 (Parkinson) and 166.79

(Garman-Klass) is statistically indistinguishable from the constant benchmark's 170.28 and 168.81. EGARCH on Bitcoin therefore provides essentially no information beyond the unconditional mean once the noise in the squared-return proxy is removed, while GARCH (78.90 and 74.03) does provide meaningful structure. The Model Confidence Set

procedure was rerun against each of the three proxies; the headline survivor sets reported in Table 9 are stable, with the LSTM-based models retained under MSE and the set expanding to include GARCH (and sometimes EGARCH) under QLIKE. The full set of MCS results across proxies is available from the corresponding author.

Table 12 Out-of-Sample MSE Across Three Volatility Proxies.

Model	Squared return	Parkinson	Garman-Klass
Panel A: BTC			
GARCH(1,1)-t	169.36	78.90	74.03
EGARCH(1,1)-t	262.51	171.74	166.78
LSTM (median)	157.78	69.26	65.10
Hybrid-A (median)	162.99	74.44	70.24
Hybrid-B (median)	157.87	69.32	65.16
Constant (uncond.)	251.35	170.28	168.81
Naive (RW)	267.38	175.47	169.76
Panel B: ETH			
GARCH(1,1)-t	730.10	314.85	293.11
EGARCH(1,1)-t	747.58	334.17	313.13
LSTM (median)	671.63	245.66	220.11
Hybrid-A (median)	815.19	427.87	416.88
Hybrid-B (median)	671.69	245.71	220.16
Constant (uncond.)	1152	789.40	787.52
Naive (RW)	1190	719.41	675.25

Notes: Out-of-sample mean squared error of each model against three alternative volatility proxies. The squared-return column reproduces the headline results from Tables 5 and 6. The Parkinson estimator is $\sigma^2_P = (\ln(H/L))^2 / (4 \ln 2)$, and the Garman-Klass estimator is $\sigma^2_{GK} = 0.5(\ln(H/L))^2 - (2 \ln 2 - 1)(\ln(C/O))^2$, both expressed in (percent squared). The relative ranking of the seven models is preserved across all three proxies for both assets, providing evidence that the headline conclusions are not driven by the noise in the squared-return proxy. $n = 1,096$.

➤ Sensitivity to the Train/Validation Split

A potential concern with any LSTM-based result is that the performance reported in Tables 5 and 6 may depend on the chosen validation share. To address this, the three LSTM-based models were re-estimated on the same pre-test sample but with two alternative splits: 75% training and 25% validation, and 80% training and 20% validation. Each alternative configuration was run with five random seeds and the median forecast was used. Table 13 reports the resulting out-of-sample MSE alongside the headline values from the 85/15 split (based on ten random seeds, as in Tables 5 and 6).

The pure LSTM and Hybrid-B are essentially invariant to the split choice. For the pure LSTM, MSE varies by less than 0.4% across the three configurations on both assets, well

within the seed-to-seed standard deviation reported in Table 10. For Hybrid-B, the corresponding variations are 0.71% on Bitcoin and 0.29% on Ethereum. Hybrid-A is more sensitive, particularly on Ethereum, where MSE varies from 736.48 at 80/20 to 850.12 at 75/25, a range of 9.66%. This pattern is consistent with the seed-instability finding reported in Section 5.5: Hybrid-A is sensitive both to random initialisation and to the validation share, whereas the other two LSTM-based models are stable on both dimensions. Importantly, the qualitative ranking is preserved at every split: the pure LSTM and Hybrid-B remain tied for the lowest MSE, and Hybrid-A trails them in every configuration for both assets. The headline conclusions are therefore not driven by the choice of an 85/15 validation share.

Table 13 Out-of-Sample Mse under Alternative Train/Validation Splits.

Model	75/25 (n=5)	80/20 (n=5)	85/15 (headline, n=10)	Max deviation
Panel A: BTC				
LSTM	158.29	157.79	157.78	0.32%
Hybrid-A	163.09	161.22	162.99	1.09%
Hybrid-B	158.99	157.89	157.87	0.71%
Panel B: ETH				
LSTM	672.49	671.38	671.63	0.13%
Hybrid-A	850.12	736.48	815.19	9.66%
Hybrid-B	673.39	671.46	671.69	0.25%

Notes: Out-of-sample MSE for each LSTM-based model under three alternative train/validation splits. The 75/25 and 80/20 configurations were estimated with five random seeds {42, 99, 271, 555, 2024} and the 85/15 configuration reproduces the headline values from Tables 5 and 6, based on ten random seeds. The maximum deviation column reports the largest absolute

percentage deviation of the 75/25 or 80/20 MSE from the 85/15 baseline. Pure LSTM and Hybrid-B are essentially invariant to the split choice, while Hybrid-A is more sensitive, consistent with the seed-instability finding in Section 5.5.

➤ Robustness to LSTM Refitting Within the Test Window

A final methodological concern is the asymmetry in how the parametric and deep learning models are updated during the test window. The GARCH and EGARCH models refit every thirty days, allowing them to adapt to new information as it arrives. The LSTM-based models in the headline results are fitted once on the pre-test data and held fixed across the entire 1,096-day test window. A reviewer could reasonably argue that this asymmetry either understates the LSTM, by denying it the chance to adapt, or overstates it, by allowing it to escape the tracking errors that the parametric models would acquire as regimes shift. To resolve this concern empirically, the three LSTM-based models were re-estimated under a rolling-refit protocol in which the LSTM is retrained from scratch every 270 days within the test window. Each refit uses an expanding training set that includes all data up to the refit date, with the most recent 15% of the training window serving as the validation set for early stopping. The architecture, lookback, and other hyperparameters are unchanged. To control for initialisation noise, three random seeds are used at each refit and the median forecast is taken. The choice of a 270-day refit interval, which produces four refits across the test window, balances comparability with the parametric models against the substantial computational cost of training the LSTM from scratch.

Table 14 compares the rolling-refit MSE, MAE, and QLIKE values to the headline fixed-LSTM results from Tables 5 and 6. The picture is clear: for five of the six asset-

and-model combinations, the rolling-refit MSE differs from the fixed-LSTM MSE by less than 1.3%, well within the seed-to-seed standard deviation reported in Table 10. The pure LSTM moves by 0.5% for Bitcoin and 0.9% for Ethereum, while Hybrid-B moves by 0.2% for Bitcoin and 0.9% for Ethereum. The qualitative ranking of the LSTM-based models is preserved under both protocols. The single notable change occurs for Hybrid-A on Ethereum, whose MSE falls from 815.19 under the fixed protocol to 734.21 under rolling refit, an improvement of 9.9%. This is consistent with the broader seed-instability and split-sensitivity pattern documented in Sections 5.5 and 5.7: Hybrid-A's loss surface is rough enough that giving the network multiple fresh starting points from different points in the test window improves average performance. Even after the rolling-refit improvement, however, Hybrid-A on Ethereum remains worse than the pure LSTM (734.21 versus 677.68), preserving the dominance ordering.

These results support a clear conclusion. The asymmetric-refit critique does not overturn the paper's findings. When the LSTM is given the same opportunity to adapt to new data as the parametric models, its performance changes by less than 1.3% in five of six cases, and the single larger change improves the worst-performing hybrid without disturbing the model ranking. The headline fixed-LSTM results reported in Tables 5 and 6 are therefore not an artefact of the chosen training protocol but reflect a genuine feature of the data.

Table 14. Out-of-sample accuracy under fixed-LSTM and rolling-refit protocols.

Model	Fixed MSE	Rolling MSE	Δ MSE	Rolling MAE	Rolling QLIKE
Panel A: BTC					
LSTM	157.78	158.58	+0.51%	5.97	2.867
Hybrid-A	162.99	161.10	-1.16%	6.55	2.838
Hybrid-B	157.87	158.19	+0.21%	6.05	2.847
Panel B: ETH					
LSTM	671.63	677.68	+0.90%	11.23	3.469
Hybrid-A	815.19	734.21	-9.93%	13.88	3.583
Hybrid-B	671.69	677.90	+0.93%	11.22	3.472

Notes: Out-of-sample accuracy comparison between the fixed-LSTM protocol used in the headline analysis (Tables 5 and 6, ten random seeds) and a rolling-refit protocol in which the LSTM is retrained from scratch every 270 days within the 1,096-day test window (three random seeds per refit, median forecast taken). The fixed-LSTM column reproduces the median MSE from Tables 5 and 6. The Δ MSE column reports the percentage change of the rolling-refit MSE relative to the fixed-LSTM MSE. The rolling-refit protocol changes MSE by less than 1.3% in five of six cases; the only exception is Hybrid-A for Ethereum, which improves by 9.9% but remains worse than the pure LSTM.

VI. DISCUSSION

The empirical results in Section 5 reveal a more nuanced picture than the headline ranking of models by mean squared error. Four findings deserve a closer reading. First, the symmetric GARCH(1,1) model dominates the asymmetric EGARCH specification on every out-of-sample metric and for both assets, while EGARCH for Bitcoin is statistically indistinguishable from a constant unconditional-variance

predictor under the range-based proxies. Second, the LSTM-based models beat the constant benchmark by 37% to 72% across proxies, confirming that the LSTM's MSE advantage reflects genuine signal rather than mean-predicting on a noisy proxy. Third, the relative ranking depends materially on the loss function but is stable across all three volatility proxies, which addresses the central concern raised by Patton (2011) about proxy-induced ranking reversals. Fourth, the two hybrid specifications produce sharply different results:

Hybrid-A is dominated by the pure LSTM and exhibits much larger seed-to-seed variation, while Hybrid-B effectively collapses to the pure LSTM rather than providing a distinct alternative.

Take the GARCH-versus-EGARCH comparison first. The full-sample EGARCH estimates reported in Table 3 show that the asymmetry parameter γ is statistically insignificant for both assets ($p = 0.083$ for Bitcoin and $p = 0.922$ for Ethereum). The information-criterion comparison in Table 4 confirms that GARCH and EGARCH are virtually indistinguishable in in-sample fit, and the out-of-sample evaluation shows that the simpler GARCH model produces substantially lower MSE for Bitcoin (a reduction from 262.51 to 169.36) and a modest improvement for Ethereum (from 747.58 to 730.10). The robustness analysis in Table 12 sharpens this picture further. Against the Parkinson and Garman-Klass proxies, the EGARCH MSE for Bitcoin (171.74 and 166.79) is essentially the same as the constant-benchmark MSE (170.28 and 168.81). In other words, once noise is partially removed from the proxy, EGARCH on Bitcoin produces no more information than predicting the unconditional mean. The standard motivation for EGARCH in equity markets, the leverage effect, does not appear to operate in the cryptocurrency sample studied here, and the practical implication is that researchers should not adopt EGARCH automatically as the parametric benchmark for cryptocurrencies; the data should be allowed to speak.

The constant-benchmark comparison is itself instructive. A common concern about LSTM-based volatility forecasts is that the network may learn to predict close to the unconditional mean of the noisy squared-return target and thereby achieve low MSE without capturing real volatility dynamics. The numerical evidence here directly addresses that concern. Against the squared-return proxy, the LSTM's MSE of 157.78 for Bitcoin is 37% below the constant benchmark's 251.35, and the LSTM's 671.63 for Ethereum is 42% below the constant's 1,152.30. The gap widens further under the cleaner range-based proxies: against Garman-Klass the LSTM beats the constant by 61% for Bitcoin and 72% for Ethereum. These margins are far too large to be attributed to chance and confirm that the LSTM is extracting genuine predictive information from the lagged return sequence, not merely producing a smoothed version of the long-run mean.

The dependence of the model ranking on the loss function deserves particular attention. The Model Confidence Set under the MSE loss retains the LSTM, Hybrid-A, and Hybrid-B for Bitcoin, and only the LSTM for Ethereum. Under the QLIKE loss, the set expands to include both parametric models for Ethereum and to include GARCH for Bitcoin, while excluding EGARCH. The contrast is sharpest for Ethereum, where the same out-of-sample data identify a set of one model under MSE and a set of four under QLIKE. This is exactly the situation Patton (2011) warns against: when the volatility proxy is noisy, point losses such as MSE can yield rankings that an asymmetric loss such as QLIKE will not corroborate. Figure 3 shows why. The LSTM forecast path hugs the unconditional mean variance with mild variation, while GARCH and EGARCH respond to volatility

regimes. The LSTM is rewarded under MSE for not overshooting in either direction, but it is penalised under QLIKE during high-volatility windows when realised variance is itself elevated and the LSTM forecast is too low. For risk-management applications, where systematic underestimation of variance is the principal concern, GARCH-type models therefore retain a useful role even when their MSE numbers look unfavourable. Crucially, the robustness check in Table 12 shows that this loss-function sensitivity is not an artefact of the squared-return proxy: rankings remain consistent under Parkinson and Garman-Klass, so the divergence between MSE-based and QLIKE-based conclusions is a genuine feature of the data and not a measurement artefact.

The divergent performance of the two hybrid specifications runs against a substantial literature. Kristjanpoller and Minutolo (2018), Seo and Kim (2020), and Zahid et al. (2022) all report that hybrid models which inject a GARCH-type variance signal as an additional input feature improve forecast accuracy relative to either component used in isolation. In the sample studied here, Hybrid-A is consistently outperformed by the pure LSTM and exhibits much larger seed-to-seed variation: the standard deviation across the ten random seeds is 3.77 for Bitcoin and 250.41 for Ethereum, compared with 0.16 and 0.32 for the pure LSTM. Two explanations are plausible. First, the EGARCH variance forecast is itself a smooth function of past returns and lagged variance, so feeding it as an additional input may not provide information beyond what the LSTM has already extracted from the raw return sequence; the added feature increases input dimensionality without adding signal. Second, EGARCH for Bitcoin is misspecified relative to a constant predictor, so injecting its variance signal can degrade rather than improve the LSTM's optimisation. Hybrid-B, which trains the LSTM on standardised residuals rather than feeding the variance forecast as an input, produces forecasts that differ from the pure LSTM by less than 0.1 MSE units in every panel, well within the seed-to-seed noise of either model. The honest interpretation is not that Hybrid-B ties the LSTM but that Hybrid-B effectively collapses to the LSTM: the EGARCH filter mostly applies a smooth scaling factor that the LSTM can absorb on its own. The takeaway is that hybrid architectures are not automatically beneficial; their value depends on whether the parametric component supplies information that the neural network would not otherwise discover.

The sub-period analysis in Table 11 confirms that these conclusions are not driven by any single phase of the test window. Volatility levels themselves vary substantially across years, with 2024 and 2025 markedly more volatile than 2023, consistent with the post-spot-ETF environment described by Babalos et al. (2025) and Whaley and Hornback (2025). However, the relative ordering of the seven models is preserved across all three years for both assets. The single anomaly, Hybrid-A's marginal lead over the pure LSTM for Bitcoin in 2024 (233.53 versus 237.24), does not survive multi-seed averaging and is dwarfed by Hybrid-A's underperformance in the other periods.

Several caveats deserve mention. First, although the use of Parkinson and Garman-Klass proxies addresses the most important concern about noise in the squared-return proxy, five-minute realised variance computed from intraday data would be a still cleaner benchmark, as Liu et al. (2015) have shown. Second, the EGARCH and GARCH specifications used here are the simplest members of their families. More flexible alternatives, including regime-switching variants (Caporale & Zekokh, 2019; Ardia et al., 2019), long-memory variants (Kaya Soyulu et al., 2020), and component-GARCH (Conrad et al., 2018), could plausibly close part of the gap to the LSTM. Third, the LSTM architecture was held fixed at two layers of thirty-two units with a twenty-two-day lookback; a more thorough hyperparameter search might raise the LSTM's performance further. The validation share, however, was checked explicitly: as Table 13 shows, the pure LSTM and Hybrid-B are essentially invariant to whether 75%, 80%, or 85% of the pre-test data is used for training, while Hybrid-A's split-sensitivity reinforces the seed-instability finding. The asymmetry between the rolling refit of the parametric models and the fixed LSTM was also addressed explicitly in Section 5.8 by re-estimating the LSTM-based models under a 270-day rolling refit; the rankings are preserved. Fourth, all models are univariate; multivariate extensions that exploit cross-asset spillovers (Walther et al., 2019; Alves et al., 2025) are left to future research. Finally, the analysis covers only Bitcoin and Ethereum; whether the same patterns hold for smaller cryptocurrencies, including those tied to specific blockchain ecosystems or stablecoin reserves, is an open empirical question.

CONCLUSION

A comprehensive comparison of seven one-day-ahead volatility forecasts for Bitcoin and Ethereum over the decade from January 2016 to December 2025 has been provided. The seven models comprise a symmetric GARCH(1,1) and an asymmetric EGARCH(1,1), both with Student-t innovations; a Long Short-Term Memory neural network with two layers of thirty-two units; a hybrid in which the EGARCH conditional variance forecast is added as an LSTM input feature (Hybrid-A); a hybrid in which the LSTM is trained on the standardised residuals of the EGARCH model (Hybrid-B); a constant unconditional-variance benchmark; and a naive random-walk benchmark. Forecasts are produced through an expanding-window rolling scheme with refit every thirty days, yielding 1,096 out-of-sample forecasts per model per asset over the test period from January 2023 to December 2025. To address the well-known sensitivity of deep learning results to random initialisation, each LSTM model is trained ten times with different seeds and the median forecast is used in evaluation. Forecast accuracy is assessed with five metrics, evaluated against the squared-return proxy and the Parkinson and Garman-Klass range-based estimators. Statistical inference is based on pairwise Diebold-Mariano tests with the Harvey, Leybourne, and Newbold (1997) small-sample correction and the Model Confidence Set procedure of Hansen, Lunde, and Nason (2011) at the ninety percent confidence level, complemented by a sub-period stability

check by calendar year and a robustness check across the three volatility proxies.

The results support five main conclusions. First, the simpler GARCH(1,1) specification consistently outperforms the asymmetric EGARCH(1,1) on both Bitcoin and Ethereum, reflecting the absence of a statistically significant leverage effect in the sample; against the range-based proxies, EGARCH for Bitcoin is statistically indistinguishable from a constant predictor. Second, the LSTM produces the most accurate point forecasts in terms of MSE, MAE, and RMSE for both assets and is the only model retained in the Model Confidence Set under MSE for Ethereum, while joining Hybrid-A and Hybrid-B in the MSE-based set for Bitcoin. The LSTM beats the constant unconditional-variance benchmark by 37% to 72% across proxies, confirming that its accuracy reflects genuine signal rather than mean prediction. Third, under the QLIKE loss the Model Confidence Set expands materially: it includes GARCH alongside the LSTM and the hybrids for Bitcoin and includes both GARCH and EGARCH for Ethereum, indicating that the parametric models remain useful when the loss function penalises variance underestimation. Fourth, this loss-function sensitivity is preserved under the cleaner Parkinson and Garman-Klass proxies, which provides direct evidence that it is a feature of the data rather than an artefact of measurement noise. Fifth, the two hybrid specifications behave very differently: Hybrid-A is dominated by the pure LSTM on every metric and exhibits substantially larger seed-to-seed variability, while Hybrid-B effectively collapses to the pure LSTM with forecasts that differ by less than one-tenth of an MSE unit. The result challenges the common assumption that combining parametric and machine learning components automatically improves forecast accuracy.

These findings carry several practical implications. For applications where point accuracy of variance forecasts is the primary concern, such as parameter inputs to a mean-variance portfolio optimiser or to a Black-Scholes-style option pricing exercise, the LSTM offers measurable improvements over both parametric benchmarks for Bitcoin and Ethereum. For risk management applications, where downside protection requires that variance not be systematically underestimated, GARCH retains a useful role and combining it with the LSTM in an ensemble may be preferable to relying on either alone. Researchers and practitioners considering hybrid architectures should test both the variance-as-feature and the residual-based constructions on their data before committing to a single design, and should report seed-to-seed variability for any LSTM-based model rather than reporting a single run.

Several extensions appear promising. The use of higher-frequency volatility proxies, including five-minute realised variance, would sharpen the rankings further. More flexible parametric benchmarks, including regime-switching and long-memory GARCH models, could narrow the gap to the LSTM. A multivariate framework that captures spillovers between Bitcoin, Ethereum, and traditional asset classes would extend the analysis to portfolio-level risk. Finally, the comparative study of hybrid architectures could be broadened to include attention-based and Transformer models, whose

application to cryptocurrency volatility remains under-explored. We leave these avenues to future work.

REFERENCES

- [1]. Alqaralleh, H., Abuhommous, A. A., & Alsaraireh, A. (2020). Forecasting cryptocurrency volatility: A comparison between GARCH-type and machine learning models. *Investment Management and Financial Innovations*, 17(4), 165–176.
- [2]. Alves, R. P., Brito, F. C. C., Mendes, P. S., & Cajueiro, D. O. (2025). Volatility transmission between cryptocurrencies and traditional financial assets: Evidence from a TVP-VAR-DY framework. *Finance Research Letters*, 65, 105827.
- [3]. Amendola, A., & Sensini, L. (2020). Conditional heteroskedastic models for cryptocurrencies volatility. *International Business Research*, 13(2), 71–82.
- [4]. Amirshahi, B., & Lahmiri, S. (2023). Hybrid deep learning and GARCH-family models for forecasting volatility of cryptocurrencies. *Machine Learning with Applications*, 12, 100465.
- [5]. Andersen, T. G., & Bollerslev, T. (1998). Answering the skeptics: Yes, standard volatility models do provide accurate forecasts. *International Economic Review*, 39(4), 885–905.
- [6]. Andersen, T. G., Bollerslev, T., Diebold, F. X., & Labys, P. (2003). Modeling and forecasting realized volatility. *Econometrica*, 71(2), 579–625.
- [7]. Aras, S. (2021). Stacking hybrid GARCH models for forecasting Bitcoin volatility. *Expert Systems with Applications*, 174, 114747.
- [8]. Aras, S., Özdemir, M. O., & Çilgin, C. (2025). Uncertainty or investor attention: Which has more impact on Bitcoin volatility? *Research in International Business and Finance*.
- [9]. Araya, H. T., Aduda, J., & Berhane, T. (2024). A hybrid GARCH and deep learning method for volatility prediction. *Journal of Applied Mathematics*, 2024, 6305525.
- [10]. Ardia, D., Bluteau, K., & Rüede, M. (2019). Regime changes in Bitcoin GARCH volatility dynamics. *Finance Research Letters*, 29, 266–271.
- [11]. Ayadi, A., Ghabri, Y., & Guesmi, K. (2025). Volatility of top market-cap cryptocurrencies during COVID-19 and the Russia–Ukraine war. *Risks*, 13(3), 57.
- [12]. Babalos, V., Bouri, E., & Gupta, R. (2025). Does the introduction of US spot Bitcoin ETFs affect spot returns and volatility of major cryptocurrencies? *The Quarterly Review of Economics and Finance*, 102, 102006.
- [13]. Bates, J. M., & Granger, C. W. J. (1969). The combination of forecasts. *Operational Research Quarterly*, 20(4), 451–468.
- [14]. Bergsli, L. Ø., Lind, A. F., Molnár, P., & Polasik, M. (2022). Forecasting volatility of Bitcoin. *Research in International Business and Finance*, 59, 101540.
- [15]. Bernardi, M., & Catania, L. (2018). The Model Confidence Set package for R. *International Journal of Computational Economics and Econometrics*, 8(2), 144–158.
- [16]. Bollerslev, T. (1986). Generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics*, 31(3), 307–327.
- [17]. Brauneis, A., & Sahiner, M. (2024). Crypto volatility forecasting: Mounting a HAR, sentiment, and machine learning horserace. *Asia-Pacific Financial Markets*.
- [18]. Buczynski, M., & Chlebus, M. (2024). GARCHNet: Value-at-risk forecasting with GARCH models based on neural networks. *Computational Economics*, 63(5), 1949–1979.
- [19]. Caporale, G. M., & Zekokh, T. (2019). Modelling volatility of cryptocurrencies using Markov-switching GARCH models. *Research in International Business and Finance*, 48, 143–155.
- [20]. Catania, L., & Grassi, S. (2022). Forecasting cryptocurrency volatility. *International Journal of Forecasting*, 38(3), 878–894.
- [21]. Cheikh, N. B., Zaied, Y. B., & Chevallier, J. (2020). Asymmetric volatility in cryptocurrency markets. *Finance Research Letters*, 35, 101293.
- [22]. Chronopoulos, I., Raftapostolos, A., & Kapetanios, G. (2024). Forecasting value-at-risk using deep neural network quantile regression. *Journal of Financial Econometrics*, 22(3), 636–669.
- [23]. Conrad, C., Custovic, A., & Ghysels, E. (2018). Long- and short-term cryptocurrency volatility components: A GARCH-MIDAS analysis. *Journal of Risk and Financial Management*, 11(2), 23.
- [24]. Corbet, S., Lucey, B., Urquhart, A., & Yarovaya, L. (2019). Cryptocurrencies as a financial asset: A systematic analysis. *International Review of Financial Analysis*, 62, 182–199.
- [25]. Corsi, F. (2009). A simple approximate long-memory model of realized volatility. *Journal of Financial Econometrics*, 7(2), 174–196.
- [26]. Cretarola, A., & Figà-Talamanca, G. (2020). Bubble regime identification in an attention-based model for Bitcoin and Ethereum price dynamics. *Economics Letters*, 191, 108831.
- [27]. D'Amato, V., Levantesi, S., & Piscopo, G. (2022). Deep learning in predicting cryptocurrency volatility. *Physica A: Statistical Mechanics and Its Applications*, 596, 127158.
- [28]. Diebold, F. X., & Mariano, R. S. (1995). Comparing predictive accuracy. *Journal of Business & Economic Statistics*, 13(3), 253–263.
- [29]. Dudek, G., Fiszeder, P., Kobus, P., & Orzeszko, W. (2024). Forecasting cryptocurrencies volatility using statistical and machine learning methods: A comparative study. *Applied Soft Computing*, 151, 111132.
- [30]. Fakhfekh, M., & Jeribi, A. (2020). Volatility dynamics of crypto-currencies' returns: Evidence from asymmetric and long memory GARCH models. *Research in International Business and Finance*, 51, 101075.
- [31]. Fiszeder, P., Małecka, M., & Molnár, P. (2024). Range-based DCC models for covariance and value-at-risk forecasting. *Journal of Empirical Finance*, 75, 101452.

- [32]. García-Medina, A., & Aguayo-Moreno, E. (2024). LSTM-GARCH hybrid model for the prediction of volatility in cryptocurrency portfolios. *Computational Economics*, 63(4), 1511–1542.
- [33]. Granger, C. W. J., & Ramanathan, R. (1984). Improved methods of combining forecasts. *Journal of Forecasting*, 3(2), 197–204.
- [34]. Hansen, P. R., & Lunde, A. (2006). Consistent ranking of volatility models. *Journal of Econometrics*, 131(1–2), 97–121.
- [35]. Hansen, P. R., Lunde, A., & Nason, J. M. (2011). The model confidence set. *Econometrica*, 79(2), 453–497.
- [36]. Harvey, D., Leybourne, S., & Newbold, P. (1997). Testing the equality of prediction mean squared errors. *International Journal of Forecasting*, 13(2), 281–291.
- [37]. Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, 9(8), 1735–1780.
- [38]. Huang, J.-Z., Huang, W., & Ni, J. (2024). Predicting Bitcoin volatility using ensemble machine learning. *Journal of Forecasting*, 43(5), 1357–1374.
- [39]. Katsiampa, P. (2017). Volatility estimation for Bitcoin: A comparison of GARCH models. *Economics Letters*, 158, 3–6.
- [40]. Katsiampa, P. (2019). Volatility co-movement between Bitcoin and Ether. *Finance Research Letters*, 30, 221–227.
- [41]. Kaya Soylu, P., Okur, M., Çatıkkaş, Ö., & Altıntig, Z. A. (2020). Long memory in the volatility of selected cryptocurrencies: Bitcoin, Ethereum and Ripple. *Journal of Risk and Financial Management*, 13(6), 107.
- [42]. Khaldi, R., El Afia, A., Chiheb, R., & Faizi, R. (2019). Forecasting of BTC volatility: Comparative study between parametric and non-parametric models. *Progress in Artificial Intelligence*, 8, 511–523.
- [43]. Köchling, G., Schmidtke, P., & Posch, P. N. (2020). Volatility forecasting accuracy for Bitcoin. *Economics Letters*, 191, 108836.
- [44]. Kristjanpoller, W., & Minutolo, M. C. (2018). A hybrid volatility forecasting framework integrating GARCH, artificial neural network, technical analysis and principal components analysis. *Expert Systems with Applications*, 109, 1–11.
- [45]. Liu, L. Y., Patton, A. J., & Sheppard, K. (2015). Does anything beat 5-minute RV? A comparison of realized measures across multiple asset classes. *Journal of Econometrics*, 187(1), 293–311.
- [46]. Maciel, L. (2021). Cryptocurrencies value-at-risk and expected shortfall: Do regime-switching volatility models improve forecasting? *International Journal of Finance & Economics*, 26(3), 4054–4072.
- [47]. Mohamed, B. T., Awad, S., & Reda, M. (2025). A comparative analysis of GARCH-family models for Bitcoin volatility. *Journal of Financial Innovation*.
- [48]. Nelson, D. B. (1991). Conditional heteroskedasticity in asset returns: A new approach. *Econometrica*, 59(2), 347–370.
- [49]. Ngunyi, A., Mundia, S., & Omari, C. (2019). Modelling volatility dynamics of cryptocurrencies using GARCH models. *Journal of Mathematical Finance*, 9(4), 591–615.
- [50]. Ormaniec, W., Pitera, M., Safarveisi, S., & Schmidt, T. (2024). Estimating value-at-risk using quantile regression and implied volatilities. *Journal of Risk*, 26(4), 23–58.
- [51]. Patton, A. J. (2011). Volatility forecast comparison using imperfect volatility proxies. *Journal of Econometrics*, 160(1), 246–256.
- [52]. Petnehazi, G. (2021). Quantile convolutional neural networks for value-at-risk forecasting. *Machine Learning with Applications*, 6, 100096.
- [53]. Phan Huy, T., Pham, K. H., Nguyen Thi Thuy, L., Nguyen Nhat, N., & Le Danh, B. (2025). Volatility dynamics in cryptocurrencies: Comparing traditional and hybrid deep learning predictive techniques. *Journal of Dynamics and Control*.
- [54]. Politis, D. N., & Romano, J. P. (1994). The stationary bootstrap. *Journal of the American Statistical Association*, 89(428), 1303–1313.
- [55]. Pratas, T. E., Ramos, F. R., & Rubio, L. (2023). Forecasting Bitcoin volatility. *Eurasian Economic Review*, 13(2), 285–305.
- [56]. Quaedvlieg, R. (2021). Multi-horizon forecast comparison. *Journal of Business & Economic Statistics*, 39(1), 40–53.
- [57]. Rodikov, G., & Antulov-Fantulin, N. (2022). Can LSTM outperform volatility-econometric models? *arXiv:2202.11581*.
- [58]. Seo, M., & Kim, G. (2020). Hybrid forecasting models based on the neural networks for the volatility of Bitcoin. *Applied Sciences*, 10(14), 4768.
- [59]. Sözen, E. (2025). Volatility dynamics of cryptocurrencies: A comparative analysis using GARCH-family models. *Future Business Journal*, 11, 166.
- [60]. Staudemeyer, R. C., & Morris, E. R. (2019). Understanding LSTM: A tutorial. *arXiv:1909.09586*.
- [61]. Trucíos, C. (2019). Forecasting Bitcoin risk measures: A robust approach. *International Journal of Forecasting*, 35(3), 836–847.
- [62]. Trucíos, C., & Taylor, J. W. (2023). A comparison of methods for forecasting value-at-risk and expected shortfall of cryptocurrencies. *Journal of Forecasting*, 42(4), 989–1007.
- [63]. Walther, T., Klein, T., & Bouri, E. (2019). Exogenous drivers of Bitcoin and cryptocurrency volatility. SSRN Working Paper No. 3192474.
- [64]. Whaley, R. E., & Hornback, A. M. (2025). Spot Bitcoin ETFs: The struggle was worth it. *Financial Analysts Journal*, 81(2), 39–50.
- [65]. Zahid, M., Iqbal, F., & Koutmos, D. (2022). Forecasting Bitcoin volatility using hybrid GARCH models with machine learning. *Risks*, 10(12), 237.
- [66]. Zhao, Y., Wu, B., & Yin, L. (2025). Integrated GARCH-GRU in financial volatility forecasting. *arXiv:2504.09380*.