DOI: https://doi.org/10.4038/sljas.v27i1.8231

Sri Lankan Journal of Applied Statistics, Vol (27-1)

Value at Risk of Stock Market Returns Using a Hidden Markov–LSTM Hybrid Model

W. A. R. De Mel*

Department of Mathematics, University of Ruhuna, Sri Lanka
*Corresponding author: ajith@maths.ruh.ac.lk

Received: 23rd September, 2025/ Revised: 16th February, 2026/ Accepted: 28th February, 2026/
Published: 31st March, 2026

ABSTRACT

Value at Risk (VaR) remains a central measure for downside market risk and is widely used in regulatory and internal risk management frameworks. Accurate VaR forecasting is challenging because equity returns exhibit volatility clustering, heavy tails, and structural changes associated with shifts between calm and turbulent market conditions. This study proposes a hybrid VaR forecasting framework that combines a Gaussian Hidden Markov Model (HMM) for latent regime identification with a Long Short Term Memory (LSTM) network for conditional volatility forecasting. The model is estimated using daily adjusted closing prices for Microsoft Corporation (MSFT) over 2016 to 2022 with a strict out of sample evaluation, where 2016 to 2019 is used for training and 2020 to 2022 is used for testing. The LSTM predicts a realized volatility proxy computed from rolling return dispersion, while the HMM regime sequence is used as an additional explanatory input to capture regime dependent dynamics. Tail risk is handled by fitting a Student t distribution to standardized residuals from the training period and constructing parametric VaR and Expected Shortfall (ES) forecasts at the 1 percent level. Out of sample backtesting using the Kupiec unconditional coverage test indicates that the observed violation frequency is close to the nominal rate and is not rejected, supporting that regime information combined with nonlinear volatility learning improves VaR calibration for equity returns.

Keywords: Forecasting; Hidden Markov Model; LSTM; Regime Switching; Backtesting; Value at Risk; Expected Shortfall

1 Introduction

Value at Risk (VaR) is a standard measure of market risk that quantifies the maximum expected loss at a given confidence level over a specified horizon. It plays a central role in internal risk management and regulatory capital frameworks (Basel Committee on Banking Supervision, 2016; Jorion, 2007). Despite its widespread adoption, accurate VaR forecasting remains challenging because financial returns exhibit volatility clustering, heavy tails, regime shifts, and nonlinear temporal dependence.

Classical conditional heteroskedasticity models such as ARCH and GARCH capture time-varying volatility through parametric structures (Engle, 1982; Bollerslev, 1986). Extensions including GJR-GARCH incorporate asymmetric responses to negative shocks (Glosten et al., 1993). While these models successfully reproduce key stylized facts of financial returns, they assume a single volatility regime evolving through a fixed functional form. Empirical evidence suggests that financial markets frequently transition between distinct volatility states, particularly during crisis episodes, which can limit the flexibility of single-regime models.

Regime-switching models provide a structured framework for capturing such structural changes. Hidden Markov Models allow the data-generating process to transition between latent states with different statistical properties (Hamilton, 1994; Rabiner, 1989). In financial applications, HMMs have been used to identify low- and high-volatility phases and to model structural breaks in return dynamics (Li and Liu, 2022). However, regime-switching models typically rely on parametric specifications that may not fully capture nonlinear temporal dependencies.

In parallel, deep learning approaches have gained attention in financial time series forecasting. Long Short-Term Memory networks are designed to model long-range dependence and nonlinear persistence while mitigating vanishing gradient problems (Hochreiter and Schmidhuber, 1997; Graves, 2013). Empirical studies show that LSTM architectures can outperform linear benchmarks in volatility and return forecasting tasks when appropriately specified (Siami-Namini et al., 2018; Livieris et al., 2020; Fischer and Krauss, 2018). Nevertheless, many deep learning VaR studies focus on direct return prediction or assume Gaussian residual structures, which may lead to misspecified tail risk estimation.

This study develops a regime-informed volatility forecasting framework for tail risk estimation. Rather than modeling raw returns, realized volatility is used as the learning target, reducing noise and improving alignment with risk measurement theory. Regime information inferred from a Hidden Markov Model is incorporated as an additional nonlinear feature within an LSTM architecture. Tail risk is then constructed using a Student-t calibration of standardized residuals estimated strictly from the training sample. This ensures methodological coherence between volatility forecasting and VaR construction.

The contribution of this paper lies in the integration of regime detection, realized volatility forecasting, heavy-tailed residual calibration, and strict out-of-sample evaluation within a single statistically consistent framework. The emphasis is not on combining two existing methods, but on constructing a coherent pipeline for tail risk estimation that addresses structural breaks, nonlinear dependence, and heavy-tailed return behavior simultaneously.

The remainder of the paper is organized as follows. Section 2 presents the methodological framework, including regime detection, volatility forecasting, and risk measure construction. Section 3 reports empirical findings and backtesting results. Section 4 concludes with implications and potential extensions.

2 Methodology

2.1 Data and return construction

Daily adjusted closing prices for Microsoft Corporation (MSFT) are obtained for the period January 2016 to December 2022. Let P_t denote the adjusted closing price at day t. Logarithmic returns are defined as

$$r_t = \log\left(\frac{P_t}{P_{t-1}}\right). \tag{1}$$

Financial returns are known to exhibit volatility clustering, heavy tails, and structural changes (Cont, 2001). To ensure reliable risk evaluation, a strict temporal split is implemented. The model is trained using data from 2016 to 2019, while the period 2020 to 2022 is reserved exclusively for out-of-sample evaluation. This separation avoids look-ahead bias and ensures that tail risk performance is assessed under genuine forecasting conditions, including the COVID-19 market stress episode.

2.2 Hidden Markov Model for regime identification

Empirical evidence suggests that financial markets transition between low- and high-volatility states (Hamilton, 1994). To capture such structural changes, a two-state Gaussian Hidden Markov Model is employed.

Let $S_t \in \{1, 2\}$ denote the latent regime at time t. Conditional on regime j, returns follow

$$r_t \mid S_t = j \sim \mathcal{N}(\mu_j, \sigma_j^2), \tag{2}$$

with transition probabilities

$$\Pr(S_t = j \mid S_{t-1} = i) = a_{ij}. \tag{3}$$

Model parameters are estimated by maximum likelihood via the Expectation–Maximization algorithm (Rabiner, 1989). The most probable regime sequence $\{\hat{S}_t\}$ is obtained through posterior decoding.

The role of the HMM in this framework is not to forecast returns directly, but to provide regime information that enriches the feature space of the nonlinear volatility model. This allows volatility dynamics to differ across latent market states.

2.3 Realized volatility as forecasting target

Directly modeling squared returns as a volatility proxy can introduce excessive noise and instability. Instead, realized volatility is constructed as a rolling measure of dispersion:

$$RV_t = \sqrt{\frac{1}{m}\sum_{k=0}^{m-1} r_{t-k}^2}, \tag{4}$$

where m denotes the rolling window length.

This specification reduces sensitivity to isolated shocks and aligns the forecasting objective with the underlying concept of conditional variance (Andersen et al., 2003; Barndorff-Nielsen and Shephard, 2002). By forecasting realized volatility rather than raw returns, the learning problem becomes smoother and more consistent with risk estimation theory.

2.4 LSTM volatility forecasting with regime inputs

Long Short-Term Memory networks are designed to capture nonlinear temporal dependence through gated memory mechanisms (Hochreiter and Schmidhuber, 1997; Graves, 2013). Let $r_t^{(s)}$ denote standardized returns obtained from the training sample. The feature vector at time t is defined as

$$x_t = \begin{bmatrix} r_t^{(s)} \\ \hat{S}_t \end{bmatrix}. \tag{5}$$

For a lookback window of length L, the input sequence is

$$X_t = (x_{t-L+1}, \ldots, x_t), \tag{6}$$

and the network produces a one-step-ahead volatility forecast $\widehat{RV}_{t+1}$.

The model is trained by minimizing mean squared error between predicted and realized volatility on the training sample. Early stopping with an internal validation split is used to mitigate overfitting. This structure allows nonlinear persistence patterns in volatility to differ across regimes identified by the HMM.

2.5 Student-t residual calibration and risk construction

Let $\hat{\sigma}_t = \widehat{RV}_t$ denote the predicted volatility. Standardized residuals are defined as

$$z_t = \frac{r_t}{\hat{\sigma}_t}. \tag{7}$$

Heavy-tailed behavior is well documented in equity returns (Cont, 2001; McNeil et al., 2015). Therefore, a Student-t distribution is fitted to $\{z_t\}$ using training data only. Let ν denote the estimated degrees of freedom.

The Value at Risk at level α is constructed as

$$VaR_{t+1}(\alpha) = t_{\alpha,\nu}\hat{\sigma}_{t+1}, \tag{8}$$

where $t_{\alpha,\nu}$ is the α quantile of the Student-t distribution.

Expected Shortfall is computed as

$$ES_{t+1}(\alpha) = -\hat{\sigma}_{t+1}\left(\frac{f_\nu(t_{\alpha,\nu})}{\alpha}\right)\left(\frac{\nu + t_{\alpha,\nu}^2}{\nu - 1}\right), \tag{9}$$

where f_ν denotes the Student-t density function (McNeil et al., 2015; Artzner et al., 1999).

Equations (4)–(10) are implemented sequentially in the empirical analysis: volatility is first forecasted, standardized residuals are computed on the training sample, the Student-t distribution is fitted, and tail quantiles are then applied to out-of-sample volatility forecasts to generate VaR and Expected Shortfall.

2.6 Backtesting and statistical validation

Let the violation indicator be defined as

$$I_{t+1} = \nVdash\{r_{t+1} < VaR_{t+1}(\alpha)\}. \tag{10}$$

The observed violation rate is

$$\hat{p} = \frac{1}{T}\sum_{t=1}^{T} I_t, \tag{11}$$

where T denotes the number of out-of-sample observations.

Unconditional coverage is assessed using the Kupiec likelihood ratio test (Kupiec, 1995). Independence and conditional coverage are evaluated using the Christoffersen test (Christoffersen, 1998). These tests jointly examine whether violations occur with the correct frequency and whether they are temporally independent, which are key regulatory requirements under Basel risk management guidelines.

In addition, rolling violation frequencies are computed to assess temporal stability of tail risk forecasts.

3 Results and Discussion

The empirical evaluation is conducted under a strict out-of-sample framework. The model is estimated using daily returns from 2016 to 2019, while all risk forecasts and statistical tests are computed exclusively over the 2020–2022 evaluation period. This interval includes the COVID-19 market disruption and the subsequent recovery phase, thereby providing a demanding stress environment for tail risk assessment. A total of 717 out-of-sample observations are used for backtesting.

3.1 Out-of-Sample Value at Risk Dynamics

Figure 1 presents realized returns together with the estimated one percent Value at Risk (VaR).

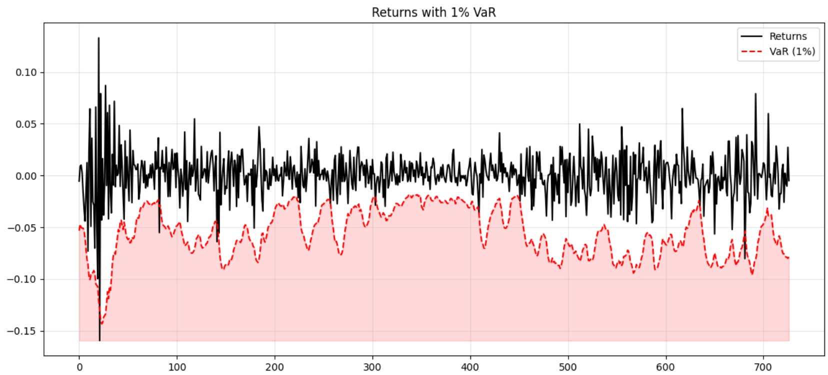

Fig. 1: Out-of-sample returns and estimated one percent Value at Risk.

The VaR forecasts adjust dynamically to changing volatility conditions. During the severe market stress observed in early 2020, the risk threshold widens substantially, reflecting elevated uncertainty. As market conditions stabilize, the VaR estimate contracts accordingly. This adaptive response indicates that realized volatility forecasting combined with regime information effectively captures structural shifts in market risk.

3.2 Backtesting Performance

Figure 2 illustrates the VaR violations observed during the out-of-sample period.

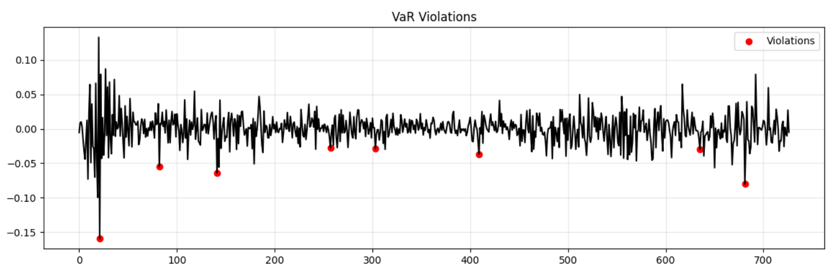

Fig. 2: VaR violations during the out-of-sample period.

Out of 717 test observations, six violations are recorded. Under the nominal one percent confidence level, approximately seven violations are expected. The observed violation rate equals 0.0084, which is closely aligned with the theoretical benchmark.

Unconditional coverage is evaluated using the Kupiec likelihood ratio test (Kupiec, 1995). The resulting p-value of 0.6514 indicates that the null hypothesis of correct coverage cannot be rejected. Independence of violations is assessed using the Christoffersen conditional coverage test (Christoffersen, 1998), yielding a p-value of 0.7501. This provides no statistical evidence of violation clustering. The absence of clustering is important, as clustered exceedances would indicate systematic underestimation of risk. Overall, the model satisfies both frequency and independence criteria commonly required for risk validation.

Figure 3 compares the expected and observed violation frequencies.

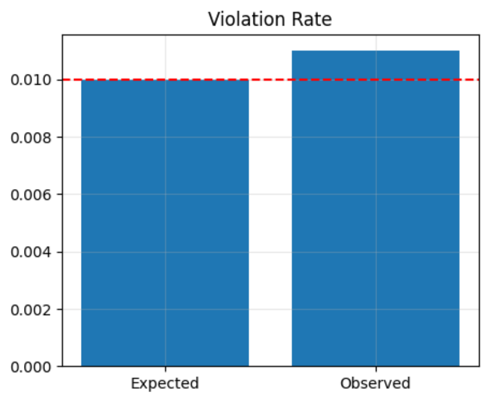

Fig. 3: Expected versus observed violation rate.

The close agreement between the expected and realized frequencies further supports the statistical adequacy of the VaR forecasts.

3.3 Expected Shortfall Behaviour

Expected Shortfall (ES) at the one percent level is shown in Figure 4.

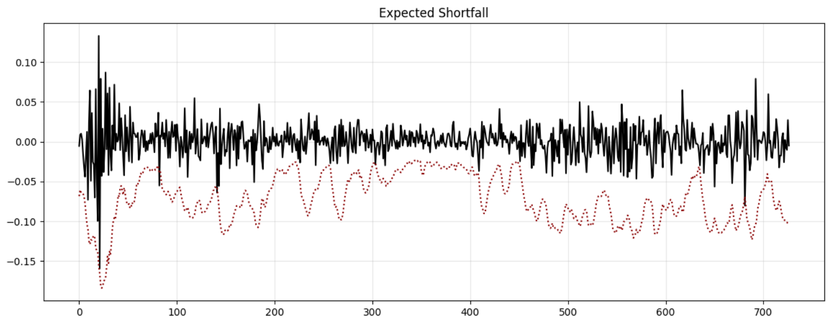

Fig. 4: Expected Shortfall at the one percent level.

Expected Shortfall increases markedly during high-volatility regimes, reflecting the magnitude of tail losses conditional on VaR exceedance. This behaviour is consistent with coherent risk measure theory (Artzner et al., 1999) and confirms that the Student-t specification captures heavy-tailed return dynamics.

3.4 Volatility Forecast Evaluation

The predicted conditional volatility series generated by the HMM–LSTM model is presented in Figure 5.

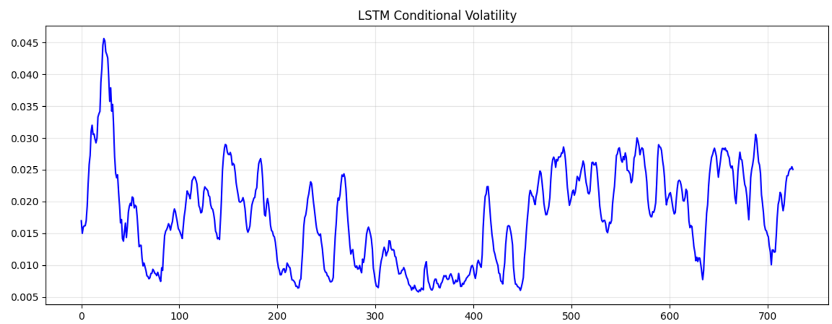

Fig. 5: Conditional volatility predicted by the HMM–LSTM model.

The volatility process exhibits clustering and persistence, with pronounced spikes during crisis periods. These characteristics are consistent with well-documented stylized facts of financial return volatility (Engle, 1982; Bollerslev, 1986).

Figure 6 compares predicted volatility with realized volatility.

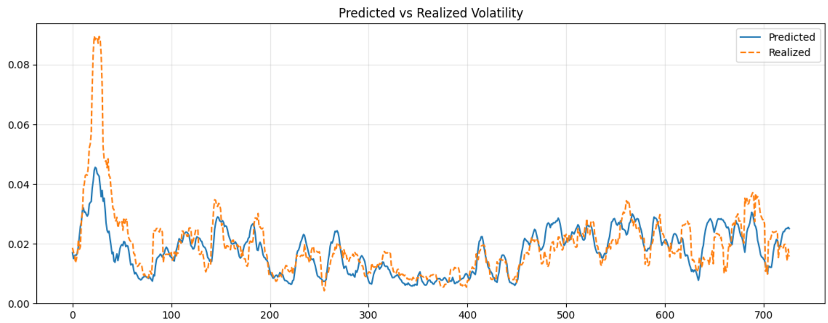

Fig. 6: Predicted versus realized volatility.

The strong co-movement between predicted and realized volatility indicates that the LSTM component effectively captures nonlinear persistence patterns. Deviations are mainly observed during abrupt volatility spikes, which are inherently difficult to forecast with precision.

3.5 Regime Identification

Figure 7 presents the volatility regimes inferred by the Hidden Markov Model.

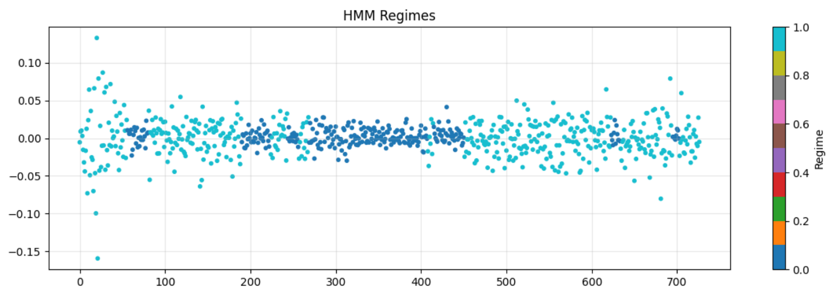

Fig. 7: Hidden Markov Model identified volatility regimes.

Two distinct regimes are identified. The high-volatility regime coincides with periods of market stress, including the COVID-19 crash, whereas the low-volatility regime corresponds to relatively stable market conditions. This economically interpretable state separation enhances nonlinear volatility forecasting and improves risk sensitivity during turbulent periods.

3.6 Residual Diagnostics

Figure 8 displays the distribution of standardized residuals.

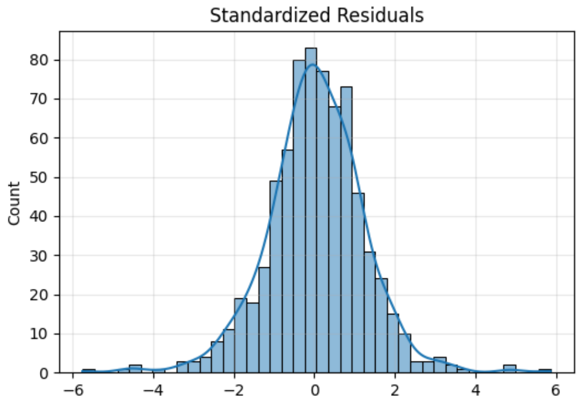

Fig. 8: Histogram of standardized residuals.

The distribution exhibits heavier tails relative to the Gaussian benchmark, supporting the Student-t assumption adopted for VaR estimation.

The quantile–quantile plot in Figure 9 further confirms this observation.

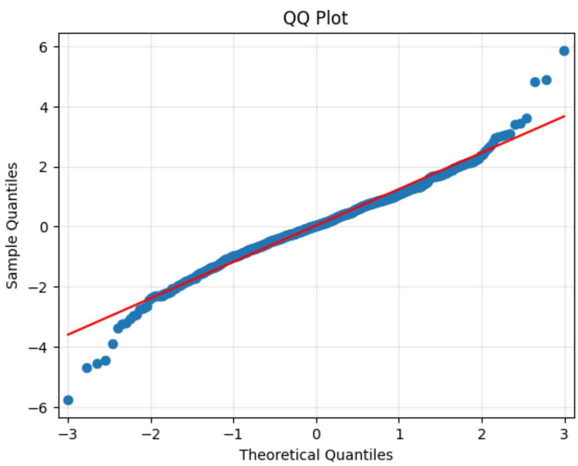

Fig. 9: Quantile–quantile plot of standardized residuals.

Noticeable deviations in the extreme quantiles provide additional justification for heavy-tailed calibration.

3.7 Rolling Stability of Tail Risk

Temporal stability of violations is examined using a rolling 100-day window, as illustrated in Figure 10.

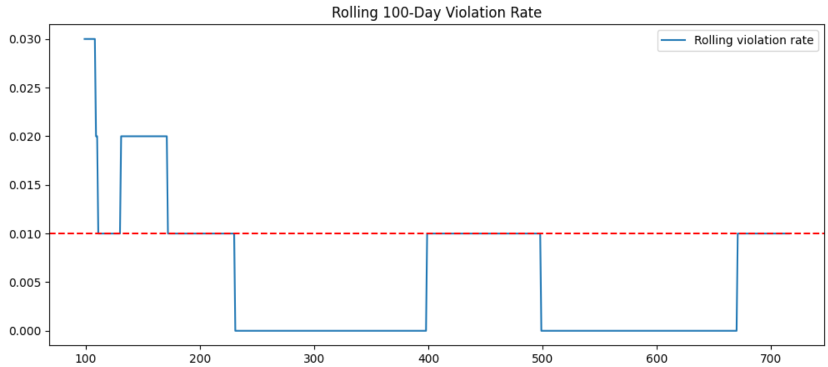

Fig. 10: Rolling 100-day violation frequency.

The rolling violation rate fluctuates around the nominal one percent benchmark without persistent exceedance. This indicates the absence of systematic underestimation of extreme losses over time.

Overall, the empirical findings demonstrate that the regime-informed realized volatility framework produces statistically valid and economically interpretable tail risk forecasts under strict out-of-sample evaluation. Both coverage and independence tests confirm the robustness of the proposed model during periods of severe market stress.

4 Conclusion

This study proposed a regime-informed realized volatility framework for tail risk forecasting under strict out-of-sample evaluation. Rather than modeling raw returns directly, the approach first extracts latent market regimes using a Hidden Markov Model and then incorporates regime information into a nonlinear Long Short-Term Memory architecture trained on a realized volatility target. One-step-ahead volatility forecasts are converted into Value at Risk and Expected Shortfall estimates using a Student-t calibration based solely on training-sample standardized residuals, thereby preserving statistical discipline and avoiding look-ahead bias.

The empirical analysis focused on Microsoft (MSFT) daily returns with a clear temporal split: 2016–2019 for model estimation and 2020–2022 for evaluation. The test period includes the COVID-19 crisis and subsequent recovery, providing a demanding stress environment. Out of 717 out-of-sample observations, six violations were observed at the one percent VaR level, compared with seven expected violations. The Kupiec unconditional coverage test yielded a p-value of 0.6514, and the Christoffersen conditional coverage test produced a p-value of 0.7501. These results indicate that the model satisfies both frequency and independence requirements for regulatory backtesting, with no evidence of systematic underestimation or clustering of extreme losses.

The findings suggest that incorporating regime information as an explicit nonlinear feature improves the stability of volatility forecasts, particularly during structural shifts. In addition, Student-t residual calibration enhances tail risk measurement relative to Gaussian assumptions, consistent with the heavy-tailed nature of financial returns documented in the risk management literature.

The contribution of this study is therefore not the mere combination of HMM and LSTM components, but the construction of a regime-informed realized volatility pipeline with disciplined tail calibration and strict post-crisis out-of-sample validation. This integrated design provides statistically coherent and economically interpretable VaR and Expected Shortfall forecasts.

Several extensions remain for future research. First, multi-asset applications can assess generalizability across equities and asset classes. Second, formal forecast comparison tests such as Diebold–Mariano statistics can be incorporated to benchmark against asymmetric GARCH-type models and other nonlinear volatility frameworks. Third, multivariate regime-informed deep learning structures may be explored for portfolio-level risk and systemic risk measurement. Finally, higher-frequency realized measures may further enhance volatility signal extraction.

Overall, the evidence indicates that regime-informed nonlinear volatility learning combined with heavy-tailed calibration offers a robust and statistically validated framework for tail risk forecasting in equity markets.

References

Andersen, T.G., Bollerslev, T., Diebold, F.X. and Labys, P. (2003) 'Modeling and Forecasting Realized Volatility', *Econometrica*, 71(2):579–625. DOI: 10.1111/1468-0262.00418.

Artzner, P., Delbaen, F., Eber, J.M. and Heath, D. (1999) 'Coherent Measures of Risk', *Mathematical Finance*, 9(3):203–228. DOI: 10.1111/1467-9965.00068.

Barndorff-Nielsen, O.E. and Shephard, N. (2002) 'Econometric Analysis of Realised Volatility and Its Use in Estimating Stochastic Volatility Models', *Journal of the Royal Statistical Society: Series B*, 64(2):253–280. DOI: 10.1111/1467-9868.00336.

Basel Committee on Banking Supervision (2016) *Minimum Capital Requirements for Market Risk*. Bank for International Settlements: Basel.

Bollerslev, T. (1986) 'Generalized Autoregressive Conditional Heteroskedasticity', *Journal of Econometrics*, 31(3):307–327. DOI: 10.1016/0304-4076(86)90063-1.

Christoffersen, P.F. (1998) 'Evaluating Interval Forecasts', *International Economic Review*, 39(4):841–862. DOI: 10.2307/2527341.

Cont, R. (2001) 'Empirical Properties of Asset Returns: Stylized Facts and Statistical Issues', *Quantitative Finance*, 1(2):223–236. DOI: 10.1080/713665670.

Engle, R.F. (1982) 'Autoregressive Conditional Heteroscedasticity with Estimates of the Variance of United Kingdom Inflation', *Econometrica*, 50(4):987–1007. DOI: 10.2307/1912773.

Fischer, T. and Krauss, C. (2018) 'Deep Learning with Long Short-Term Memory Networks for Financial Market Predictions', *European Journal of Operational Research*, 270(2):654–669. DOI: 10.1016/j.ejor.2017.11.054.

Glosten, L.R., Jagannathan, R. and Runkle, D.E. (1993) 'On the Relation between the Expected Value and the Volatility of the Nominal Excess Return on Stocks', *The Journal of Finance*, 48(5):1779–1801. DOI: 10.1111/j.1540-6261.1993.tb05128.x.

Graves, A. (2013) 'Generating Sequences with Recurrent Neural Networks', *arXiv preprint arXiv:1308.0850*.

Hamilton, J.D. (1994) *Time Series Analysis*. Princeton University Press: Princeton, NJ.

Hochreiter, S. and Schmidhuber, J. (1997) 'Long Short-Term Memory', *Neural Computation*, 9(8):1735–1780. DOI: 10.1162/neco.1997.9.8.1735.

Jorion, P. (2007) *Value at Risk: The New Benchmark for Managing Financial Risk*. 3rd edn. McGraw-Hill: New York.

Kupiec, P.H. (1995) 'Techniques for Verifying the Accuracy of Risk Measurement Models', *The Journal of Derivatives*, 3(2):73–84. DOI: 10.3905/jod.1995.407942.

Li, X. and Liu, Y. (2022) 'Regime-Switching Models in Finance: A Review and New Insights', *Journal of Risk and Financial Management*, 15(2):60. DOI: 10.3390/jrfm15020060.

Livieris, I.E., Pintelas, E. and Pintelas, P. (2020) 'A CNN-LSTM Model for Gold Price Time-Series Forecasting', *Neural Computing and Applications*, 32:17351–17360. DOI: 10.1007/s00521-020-05004-5.

McNeil, A.J., Frey, R. and Embrechts, P. (2015) *Quantitative Risk Management: Concepts, Techniques and Tools*. Revised edn. Princeton University Press: Princeton, NJ.

Rabiner, L.R. (1989) ’A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition’, *Proceedings of the IEEE*, 77(2):257–286. DOI: 10.1109/5.18626.

Siami-Namini, S., Tavakoli, N. and Siami Namin, A. (2018) ’A Comparison of ARIMA and LSTM in Forecasting Time Series’, *Proceedings of the 17th IEEE International Conference on Machine Learning and Applications (ICMLA)*, 1394–1401. DOI: 10.1109/ICMLA.2018.00227.