Available online at http://scik.org
Commun. Math. Biol. Neurosci. 2026, 2026:48
https://doi.org/10.28919/cmbn/9834
ISSN: 2052-2541

A HYBRID MULTI-GARCH TRANSFORMER MODEL FOR STOCK PRICE VOLATILITY FORECASTING

TIRHAS TESFAY GEBRESLASE[1,3,*], KILAI MUTUA[2], YEMANE HAILU FISSUH[3], TSGAB GEBRECHERKOS GIRMAY[3]

[1]Department of Mathematics, Institute for Basic Sciences, Technology and Innovation, Pan African University, 62000-00200, Nairobi, Kenya

[2]Department of Pure and Applied Sciences, Kirinyaga University, 143-10300 Kerugoya, Kenya

[3]Department of Statistics, Aksum University, Aksum 1010, Ethiopia

Abstract. Accurate forecasting of financial market volatility remains a central challenge due to nonlinear dynamics, volatility clustering, and regime-dependent behavior in financial time series. While econometric volatility models capture conditional heteroskedasticity, deep learning architectures provide flexible nonlinear modeling capabilities. However, existing approaches often treat these paradigms independently or combine them post hoc, limiting their ability to jointly exploit statistical structure and representation learning. This study addresses this gap by proposing a hybrid architecture that integrates GARCH-family models directly within an attention-based Transformer through representation-level learning. We develop a Multi-GARCH Transformer architecture that embeds conditional variance estimates from multiple GARCH-family models into the input representation of a Transformer encoder. This hybrid architecture functions as a heterogeneous ensemble in which complementary volatility dynamics are integrated at the feature level and adaptively weighted through self-attention. The analysis uses daily stock returns of Safaricom from 2015 to 2024.The study combines descriptive statistics, econometric modeling, and transformer-based hybrid forecasting. Model performance is evaluated using quasi-likelihood

*Corresponding author
E-mail address: tirhasgeb2023@gmail.com
Received February 20, 2026

(QLIKE), RMSE, and MAE, alongside statistical comparison using the Diebold-Mariano test. Empirical findings confirm that the Multi-GARCH-Transformer ensemble achieves superior out-of-sample volatility forecasting performance by embedding complementary econometric signals within an attention-based architecture. This hybrid model delivers the strongest performance under QLIKE and exhibits statistically significant improvements validated through Diebold-Mariano testing. The observed training behavior indicates more stable convergence and stronger generalization than non-hybrid architectures, reinforcing the effectiveness of representation-level ensemble learning. The results highlight the advantages of integrating econometric theory with attention-based deep learning through feature-level hybridization. By jointly modeling structured volatility dynamics and nonlinear temporal dependencies, the approach improves forecasting robustness, interpretability, and predictive accuracy. This study contributes to the intersection of financial econometrics and deep learning by introducing a hybrid ensemble modeling strategy for complex and evolving market environments.

Keywords: volatility forecasting; hybrid ensemble learning; GARCH models; transformer architecture; financial time series.

2020 AMS Subject Classification: 91G70.

1. INTRODUCTION

Accurate modelling and forecasting of financial market volatility remain challenging in quantitative finance, with critical implications for risk management, derivative pricing, portfolio optimisation and financial stability assessment. Financial return series exhibit well-documented stylised characteristics, including heavy-tailed distributions, volatility clustering, leverage effects and nonlinear temporal dependence [1, 12, 15]. These empirical properties complicate predictive modelling and require methodologies capable of capturing both structured statistical dynamics and complex nonlinear behaviour. Developing forecasting frameworks that combine theoretical foundations with flexible representation learning, therefore, represents a key research frontier.

Traditional econometric volatility models, particularly ARCH and GARCH-family, have provided foundational tools for modelling conditional heteroskedasticity in financial returns. The ARCH model introduced by Engle [1] and its generalization by Bollerslev [12] established statistically grounded approaches for modelling time-varying variance dynamics. Later extensions, including EGARCH [13] and GJR-GARCH [14], incorporated asymmetric responses to shocks and leverage effects which improved representation of observed market behaviour.

Therefore, although these models possess strong theoretical foundations and interpretability, their parametric structure constrains their ability to represent nonlinear dynamics, regime shifts and long-range dependencies in evolving financial markets.

Recent machine learning advancements have introduced powerful data-driven approaches capable of modelling complex temporal structures. Deep learning architectures suchas recurrent neural networks and long short-term memory models, have demonstrated promising performance in financial forecasting tasks [16, 17]. Specifically, attention-based architectures such as Transformers have transformed sequence modelling by enabling efficient learning of long-range dependencies through self-attention mechanisms [8]. The Transformer variants designed for long time-series forecasting, including Informer [9], further underscore the potential of attention-based learning for complex temporal prediction problems. However, these approaches often lack explicit incorporation of domain-specific statistical knowledge, potentially reducing interpretability and robustness when applied to noisy and non-stationary financial environments.

Consquently, Hybrid ensemble modelling strategies have emerged as a promising direction for combining econometric theory with deep learning flexibility [10, 4]. Existing hybrid models typically integrate models sequentially or combine outputs after independent estimation which limits the ability of deep architectures to fully exploit structured statistical information during representation learning. Emerging market financial series often exhibit heightened structural instability, liquidity constraints and evolving volatility regimes which motivates hybrid approaches that combine statistical robustness with flexible representation learning. Consequently, an important research gap remains in developing hybrid architectures that embed econometric volatility signals directly into deep learning pipelines in a principled and integrated manner.

Therefore, this study proposes a hybrid Multi-GARCH-Informed Transformer architecture for volatility forecasting that explicitly addresses this gap.This ensemble architecture embeds conditional variance estimates from multiple GARCH-family models directly into the input representation of a Transformer encoder. This feature-level integration enables the attention mechanism to jointly learn persistence, asymmetry and nonlinear temporal dependencies, allowing structured econometric volatility dynamics and flexible deep learning representations to interact during model training.

The main objective of this study was to develop and empirically evaluate a principled hybrid Multi-GARCH-Transformer architecture that integrates econometric volatility modelling with attention-based deep learning in order to improve the accuracy, robustness and interpretability of financial volatility prediction, using Safaricom daily stock returns as Data-driven analysis case study.

The study makes three key contributions. First, it proposes a hybrid Multi-GARCH-Transformer architecture that embeds econometric volatility signals directly within a Transformer, allowing structured statistical information to guide representation learning rather than combining outputs post hoc. Second, it provides strong out-of-sample evidence of statistically significant improvements in volatility forecasting within an emerging market context. Third, it demonstrates how integrating econometric modelling with attention-based learning enhances robustness, interpretability, and predictive stability in financial time-series forecasting.

2. METHODS

This study developed a hybrid Multi-GARCH-Transformer ensemble volatility forecasting architecture that integrates econometric volatility modelling with deep learning through an encoder-only Transformer architecture. This methodology combines the statistical structure of multiple GARCH-family models with attention-based representation ensemble learning which enables joint modelling of persistence, asymmetry and nonlinear temporal dependencies. This model architecture consists of data transformation, econometric volatility estimation, ensemble-based hybrid feature construction, Transformer-based sequence modelling and systematic evaluation with statistical testing.

2.1. Data source and computational environment. Daily closing price data for Safaricom PLC were retrieved from publicly available financial market databases, specifically Market-Watch (https://www.marketwatch.com), covering the full empirical study period(2015-2024). Logarithmic returns were computed from adjusted closing prices to ensure temporal consistency, scale invariance and comparability in volatility modelling. Missing values were imputed using cubic spline interpolation, where for a series of points (x_i, y_i) the cubic spline $S(x)$ on each interval $[x_i, x_{i+1}]$ is defined as

$$S_i(x) = a_i + b_i(x - x_i) + c_i(x - x_i)^2 + d_i(x - x_i)^3, \quad i = 1, \ldots, n-1. \tag{1}$$

All analyses were conducted using Python version 3.10. Econometric volatility models were estimated using `arch` and `statsmodels`, while Transformer-based architectures were implemented in PyTorch. Data preprocessing, feature scaling, and transformations utilised `NumPy`, `pandas`, and `scikit-learn`. A strict chronological split was applied to preserve temporal structure and avoid information leakage. Reproducibility was ensured using fixed random seeds. Hyperparameters (learning rate, window size, attention heads and encoder depth) were selected through validation-based optimisation. Standardisation (zero mean, unit variance) was applied to all input features:

$$x_{\text{std}} = \frac{x - \mu}{\sigma}, \tag{2}$$

where μ and σ are the mean and standard deviation estimated from the training set.

2.2. Problem formulation and data transformation. Let P_t denote the daily closing price at time t. Logarithmic returns are defined as

$$r_t = \log(P_t) - \log(P_{t-1}), \tag{3}$$

which ensures scale invariance and stabilizes variance. The forecasting objective is the conditional variance

$$\sigma_t^2 = \mathbb{E}[r_t^2 \mid \mathscr{F}_{t-1}], \tag{4}$$

where $\mathscr{F}_{t-1}$ denotes the information set available at time $t-1$. Realised volatility (RV) serves as a proxy for latent volatility and its logarithm is used as the prediction target to improve numerical stability during neural network training. The explicit one-step-ahead target is

$$y_t = \log(\text{RV}_{t+1} + \varepsilon), \quad \varepsilon = 10^{-12}. \tag{5}$$

2.3. Econometric volatility modelling. Three GARCH-family models capture complementary volatility dynamics, differently.

GARCH(1,1). The standard GARCH Model [1, 12] is defined as

$$\sigma_t^2 = \omega + \alpha\varepsilon_{t-1}^2 + \beta\sigma_{t-1}^2, \tag{6}$$

where ε_t denotes innovation terms.

EGARCH(1,1). The exponential GARCH model [13] captures asymmetric responses through

$$\log\sigma_t^2 = \omega + \beta\log\sigma_{t-1}^2 + \alpha\left(\frac{|\varepsilon_{t-1}|}{\sigma_{t-1}} - \sqrt{\frac{2}{\pi}}\right) + \gamma\frac{\varepsilon_{t-1}}{\sigma_{t-1}}, \tag{7}$$

where the term $\mathbb{E}[|\varepsilon_{t-1}/\sigma_{t-1}|] = \sqrt{2/\pi}$ under normality.

GJR-GARCH(1,1). The GJR-GARCH process [14] models leverage effects:

$$\sigma_t^2 = \omega + \alpha\varepsilon_{t-1}^2 + \gamma I_{\{\varepsilon_{t-1}<0\}}\varepsilon_{t-1}^2 + \beta\sigma_{t-1}^2, \tag{8}$$

where $I_{\{\cdot\}}$ is an indicator function which captures asymmetric responses to negative shocks. All parameters are estimated exclusively on the training set to prevent information leakage. One-step-ahead conditional variance forecasts are generated recursively and used as structured inputs.

2.4. Hybrid feature construction. The proposed hybrid Multi-GARCH-Transformer architecture adopts an ensemble-learning perspective by combining multiple heterogeneous econometric volatility estimators into a unified feature representation. Rather than aggregating forecasts through traditional model averaging or voting strategies [11], the ensemble is constructed at the feature level, where conditional variance estimates from complementary GARCH-family models serve as structured inputs to the Transformer encoder. This representation-level ensemble design allows the neural architecture to learn adaptive weighting and interaction patterns among ensemble members during training, effectively integrating diverse statistical representations of volatility dynamics. Embedding structured domain knowledge directly within representation learning has been shown to improve generalisation and interpretability by introducing inductive biases that constrain the hypothesis space [3, 6]. Similar hybrid econometric-deep learning approaches demonstrate that integrating model based statistical signals with neural architectures enhances forecasting robustness in financial time series [10]. Let

$$v_t = \left[\sigma_{t,\mathrm{GARCH}}^2, \sigma_{t,\mathrm{EGARCH}}^2, \sigma_{t,\mathrm{GJR}}^2\right]^\top, \tag{9}$$

denote econometric volatility signals. Hybrid feature vectors are defined as

$$z_t = [r_t, v_t], \tag{10}$$

allowing the Transformer to jointly learn from raw returns and structured econometric volatility information.

2.5. Transformer encoder architecture. An encoder-only Transformer architecture [8] processes hybrid sequences (Figure 1). Inputs are embedded, augmented with positional encoding, and passed through stacked encoder blocks comprising multi-head self-attention, feed-forward layers, residual connections and layer normalization. The sinusoidal positional encoding [8] for position p and dimension i is given by

$$\mathrm{PE}(p,2i) = \sin\left(\frac{p}{10000^{2i/d_{\text{model}}}}\right), \quad \mathrm{PE}(p,2i+1) = \cos\left(\frac{p}{10000^{2i/d_{\text{model}}}}\right). \tag{11}$$

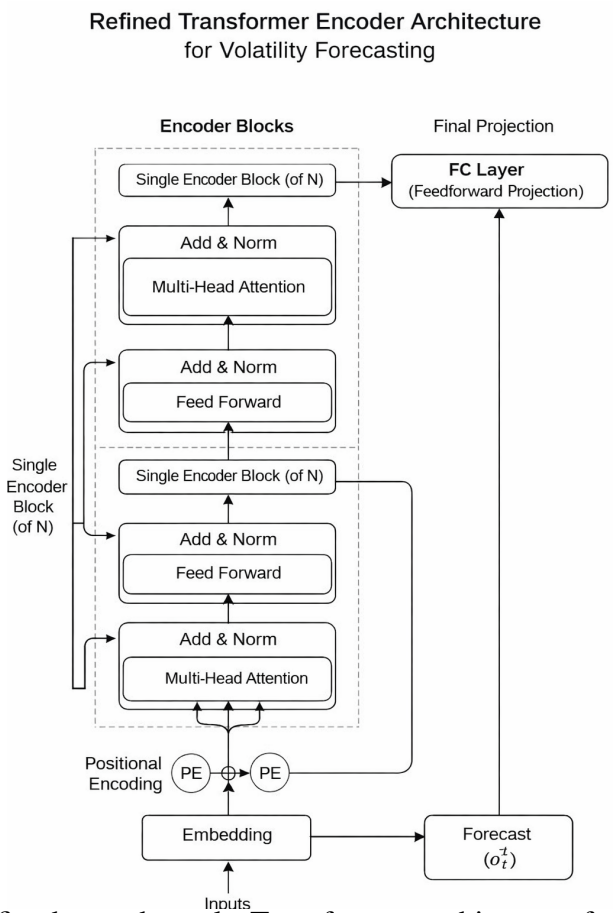

FIGURE 1. Refined encoder-only Transformer architecture for volatility forecasting with hybrid inputs.

Given window length L, the input sequence is

$$X_t = [z_{t-L+1}, \ldots, z_t]. \tag{12}$$

Self-attention is computed as

$$\text{Attention}(Q,K,V) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right)V. \tag{13}$$

The encoder output is mapped to a one-step-ahead volatility forecast:

$$\hat{y}_t = f_\theta(X_t), \tag{14}$$

where $f_\theta(\cdot)$ denotes the Transformer parameterized by θ.

2.6. Model taxonomy and experimental configurations. To systematically evaluate the contribution of ensemble components and hybrid feature integration, twelve model variants are defined (Table 1).

TABLE 1. Model taxonomy and experimental configurations.

Model ID	Model Name	Input Features	Description
M1	Baseline Transformer (Standard)	r_t	Returns-only Transformer
M2	Baseline Transformer (Residual)	r_t	Returns-only with residual correction
M3	GARCH-Informed (Standard)	$[r_t, \sigma^2_{GARCH}]$	Hybrid with GARCH input
M4	GARCH-Informed (Residual)	$[r_t, \sigma^2_{GARCH}]$	Residual hybrid
M5	EGARCH-Informed (Standard)	$[r_t, \sigma^2_{EGARCH}]$	Asymmetric hybrid
M6	EGARCH-Informed (Residual)	$[r_t, \sigma^2_{EGARCH}]$	Residual asymmetric hybrid
M7	GJR-Informed (Standard)	$[r_t, \sigma^2_{GJR}]$	Leverage-aware hybrid
M8	GJR-Informed (Residual)	$[r_t, \sigma^2_{GJR}]$	Residual leverage-aware hybrid
M9	Multi-GARCH-Only (Standard)	v_t	Volatility-only Transformer
M10	Multi-GARCH-Only (Residual)	v_t	Residual volatility-only model
M11	Multi-GARCH-Informed (Standard)	$[r_t, v_t]$	Full hybrid architecture
M12	Multi-GARCH-Informed (Residual)	$[r_t, v_t]$	Residual-enhanced full hybrid

Residual variants incorporate linear econometric correction:

$$\hat{y}_{t+1} = f_\theta(X_t) + w^\top z_t, \tag{15}$$

Prediction-level ensembles are also constructed from the single-hybrid models (GARCH, EGARCH, GJR). The simple average ensemble is

$$\hat{y}_t^{\text{avg}} = \frac{1}{3}\left(\hat{y}_t^{\text{GARCH}} + \hat{y}_t^{\text{EGARCH}} + \hat{y}_t^{\text{GJR}}\right), \tag{16}$$

and the validation-weighted ensemble solves

$$\mathbf{w}^* = \arg\min_{\mathbf{w}\geq 0} \sum_{t\in\text{val}} \left(y_t - \sum_{k=1}^{3} w_k \hat{y}_t^{(k)}\right)^2, \quad \text{s.t.} \sum_{k=1}^{3} w_k = 1, \tag{17}$$

with weights applied as $\hat{y}_t^{\text{weighted}} = \sum_k w_k^* \hat{y}_t^{(k)}$.

2.7. Evaluation metrics and statistical testing. Forecast accuracy is evaluated using RMSE, MAE and the quasi-likelihood loss (QLIKE):

$$\text{QLIKE} = \frac{1}{T}\sum_{t=1}^{T}\left(\frac{RV_t}{\hat{\sigma}_t^2} - \log\frac{RV_t}{\hat{\sigma}_t^2} - 1\right). \tag{18}$$

Statistical significance of performance differences is assessed using the Diebold-Mariano test [18]:

$$DM = \frac{\bar{d}}{\sqrt{\hat{\sigma}_d^2/T}}, \tag{19}$$

where $\bar{d}$ denotes the mean loss differential, and $\hat{\sigma}_d^2$ is a heteroskedasticity and autocorrelation consistent (HAC) variance estimator. This predictive architecture integrates econometric volatility estimation, hybrid ensemble feature construction and Transformer-based sequence modelling within a unified approach designed to jointly leverage statistical structure and deep representation learning for volatility forecasting.

3. Results and Discussion

This section presents Experimental results of modelling and forecasting the volatility of Safaricom daily stock returns using econometric benchmark models and hybrid Transformer architectures. Results are organized to evaluate statistical properties of the data, adequacy of econometric volatility models, comparative out-of-sample forecasting performance and implications for volatility-driven risk assessment. All model variants are trained using identical data

splits, preprocessing procedures and optimisation settings to ensure fair and unbiased comparison. Importantly, the proposed ensemble architecture operates as a heterogeneous ensemble learning system in which multiple econometric volatility estimators provide complementary and different structured signals that are jointly integrated within the Transformer representation space through representation-level ensemble learning. Unlike classical ensemble aggregation methods, weighting of component signals is learned implicitly via the attention mechanism whcih enables adaptive integration of heterogeneous volatility perspectives across time.

3.1. Descriptive statistics and stylised facts. Table 2 shows summarises of distributional properties of daily closing prices. The sample contains 2,498 observations with substantial dispersion, reflecting evolving market conditions and structural variability typical of emerging equity markets. The Positive skewness indicates asymmetric upward price movements, while negative kurtosis suggests deviations from Gaussian assumptions and potential regime-dependent behaviour. These characteristics motivate modelling approaches capable of capturing nonlinear and time-varying volatility dynamics.

TABLE 2. Descriptive statistics of daily closing prices

Series	Count	Mean	Std	Min	25%	50%	75%	Max	Skewness	Kurtosis
Price	2498	24.3564	8.0438	11.65	16.85	24.25	29.25	44.95	0.5187	-0.6181

Table 3 Daily log returns exhibit heavy tails, excess kurtosis and strong rejection of normality (JB = 1972.43, $p < 0.001$), confirming canonical stylised facts of financial time series including volatility clustering and conditional heteroskedasticity [1, 12, 15]. These methods justify the joint use of econometric volatility models which encode persistence and asymmetry, together with flexible attention-based architectures capable of modelling nonlinear and regime-dependent variance dynamics.

TABLE 3. Descriptive statistics of daily log returns

Series	Count	Mean	Std	Min	25%	50%	75%	Max	Skewness	Kurtosis	JB	p-value
Return	2497	0.000075	0.01702	-0.10385	-0.00770	0.00000	0.00777	0.09531	0.0625	4.3634	1972.43	0.0000

Figure 2 presents the daily closing price series and reveals pronounced non-stationary behaviour characterised by trending dynamics and evolving variance over time. The absence of mean-reverting behaviour in levels suggests violation of weak stationarity assumptions, motivating analysis in log-return space where differencing stabilises statistical properties and removes stochastic trends.

Figure 3 provides an alternative visualisation of the price dynamics through a time-series representation of daily closing prices, further highlighting structural changes and persistence in levels. The apparent temporal dependence in price levels reinforces the need for transformation into returns prior to modelling, as direct modelling in levels may lead to spurious inference.

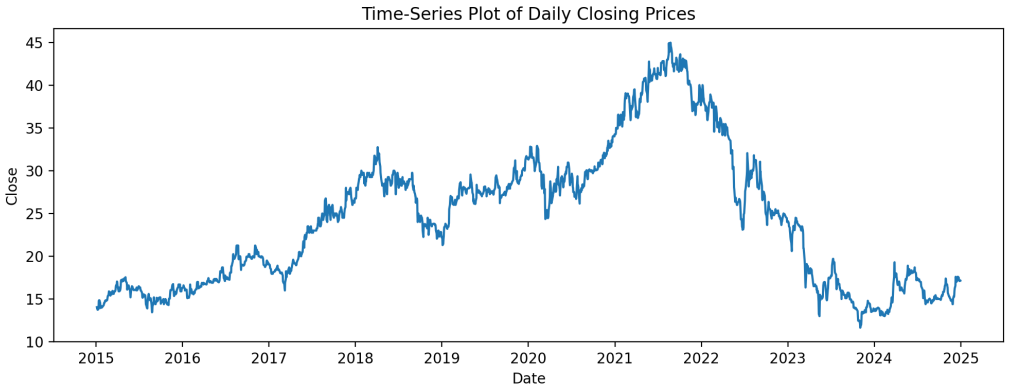

FIGURE 2. Daily closing price series illustrating non-stationary behaviour.

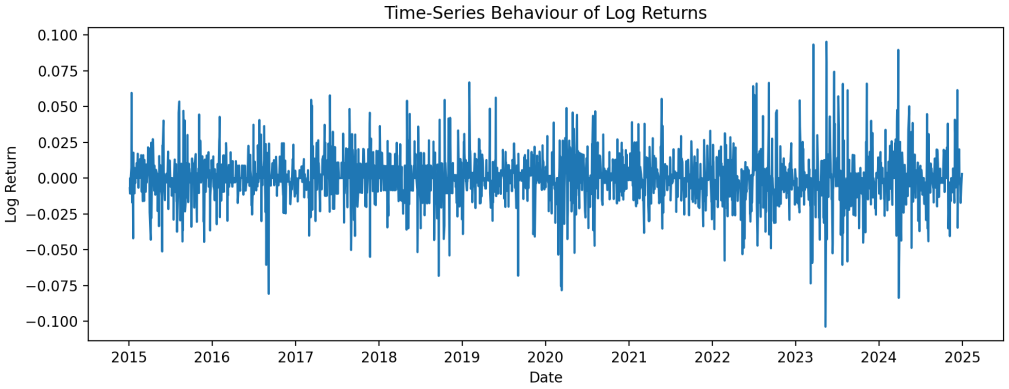

FIGURE 3. Time-series plot of daily closing prices.

Autocorrelation analysis of higher-order moments, as illustrated in Figure 4, reveals significant and slowly decaying dependence in squared returns. This behaviour provides strong empirical evidence of volatility clustering and persistent conditional heteroskedasticity, supporting the inclusion of GARCH-family models as structured inductive biases within the hybrid modelling framework. Such models explicitly capture variance persistence, while attention-based architectures allow flexible modelling of nonlinear and regime-dependent volatility dynamics.

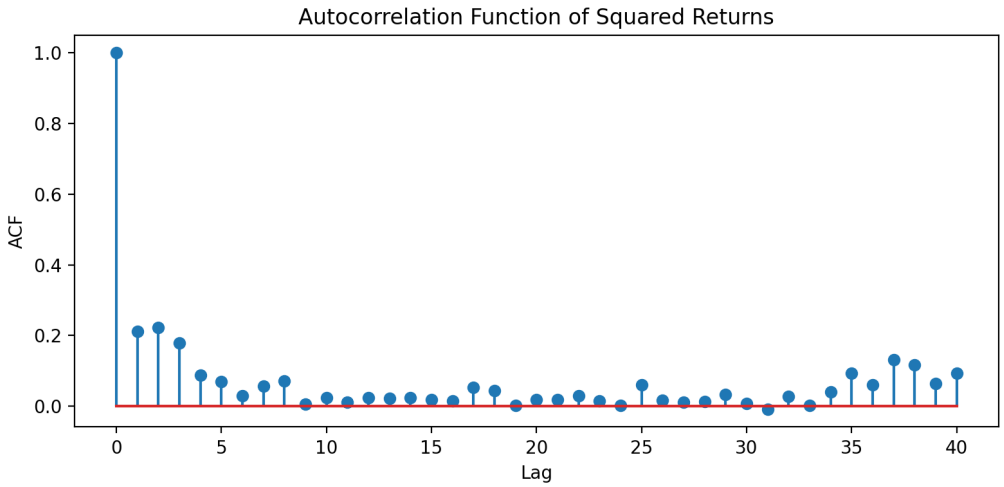

FIGURE 4. Autocorrelation function of squared returns showing volatility clustering.

3.2. Econometric benchmark evaluation. The lag-order selection guided by the Bayesian Information Criterion consistently favours parsimonious $(1,1)$ specifications across all GARCH-family models, reflecting the well established empirical finding that low order volatility dynamics often capture the dominant persistence structure in financial returns [12]. Within this modelling architecture, the standard GARCH model provides a baseline representation of conditional variance dynamics through symmetric responses to shocks, capturing volatility persistence via lagged squared innovations and lagged conditional variance. The EGARCH specification extends this structure by modelling volatility in logarithmic form and introducing asymmetric magnitude sensitivity, thereby allowing volatility responses to depend on both the size and sign of past shocks. In contrast, the GJR-GARCH model explicitly operationalises leverage effects, enabling negative shocks to exert disproportionately stronger impacts on future volatility relative to positive shocks of comparable magnitude [14].

Estimated parameters obtained in this study reveal pronounced persistence across all specifications, indicating gradual decay of volatility shocks and confirming the presence of long-memory dynamics in conditional variance processes. These findings are consistent with the stylised behaviour of equity return volatility and reinforce the role of GARCH-family models as structurally grounded econometric benchmarks.

From a representation-learning perspective, the econometric models are interpreted not merely as competing alternatives but as components of a heterogeneous ensemble of structured volatility representations. Each model contributes a distinct inductive bias reflecting different aspects of volatility dynamics: persistence and mean-reverting conditional variance captured by GARCH, asymmetric magnitude responses encoded by EGARCH, and leverage-driven asymmetry formalised by GJR-GARCH. Treating these models as complementary signal generators extends hybrid forecasting paradigms that integrate statistical modelling with neural representation learning rather than relying on single-model assumptions [7].

Embedding these complementary econometric structures directly into the learning pipeline constrains the hypothesis space toward financially meaningful temporal patterns, thereby improving model generalisation while preserving economic interpretability. Such representation-level integration aligns with modern advances in deep learning theory, which emphasise the role of domain-specific inductive biases in guiding representation learning and stabilising optimisation in complex sequential forecasting tasks [3, 10].

3.3. Out-of-sample forecasting performance. Table 4 shows forecasting accuracy across model variants. The Multi-GARCH-Informed Transformer (Standard) achieves the lowest QLIKE value (-0.719), indicating reduced variance mis-specification and superior conditional variance forecasting performance. Improvements are consistent across multiple metrics, suggesting robust performance rather than optimisation toward a single loss function. The ranking hierarchy further reveals that multi-source hybrid transformer outperform both single-source hybrids and deep learning models relying solely on returns which supportes the interpretation of ensemble diversity contributes meaningful complementary information.

TABLE 4. Out-of-sample forecasting performance of baseline and hybrid Transformer-GARCH models

Model	$RMSE_{RV}$	MAE_{RV}	$RMSE_{\log RV}$	$MAE_{\log RV}$	QLIKE	Rank
Multi-GARCH-Informed Transformer (Standard)	0.000903	0.000346	7.580	4.480	-0.719	1
Baseline Transformer (Returns-Based, Standard)	0.000906	0.000351	7.254	4.743	14.752	2
Multi-GARCH-Only Transformer (Standard)	0.000900	0.000344	7.374	4.743	15.641	3
Multi-GARCH-Informed Transformer (Residual)	0.000901	0.000348	7.332	4.801	36.457	4
GJR-GARCH-Informed Transformer (Residual)	0.000893	0.000345	7.181	5.093	37.412	5
GJR-GARCH-Informed Transformer (Standard)	0.000898	0.000352	7.205	5.115	39.089	6
EGARCH-Informed Transformer (Standard)	0.000883	0.000349	7.215	5.077	39.347	7
Baseline Transformer (Returns-Based, Residual)	0.002099	0.000483	7.209	5.087	42.470	8
EGARCH-Informed Transformer (Residual)	0.000899	0.000351	7.242	5.123	42.528	9
GARCH-Informed Transformer (Standard)	0.000898	0.000347	7.258	5.106	45.321	10
Multi-GARCH-Only Transformer (Residual)	0.000903	0.000348	7.280	5.204	50.953	11
GARCH-Informed Transformer (Residual)	0.000899	0.000349	7.232	5.376	72.422	12

The superior performance can be interpreted through complementary inductive biases introduced by representation-level ensemble learning. Attention mechanisms dynamically reconcile persistence signals, asymmetric responses and leverage effects which enabls adaptive regime-sensitive forecasting. Transformer architectures have been shown to capture long-range dependencies in time-series forecasting tasks [8, 9], while hybrid econometric-deep learning models

enhance robustness by incorporating domain-specific structure [4]. Moreover, the effectiveness of QLIKE evaluation aligns with volatility forecast comparison theory [2].

Figure 5 demonstrates smoother and more stable convergence for ensemble-informed hybrid models, consistent with reduced optimisation variance when structured econometric signals guide representation learning.

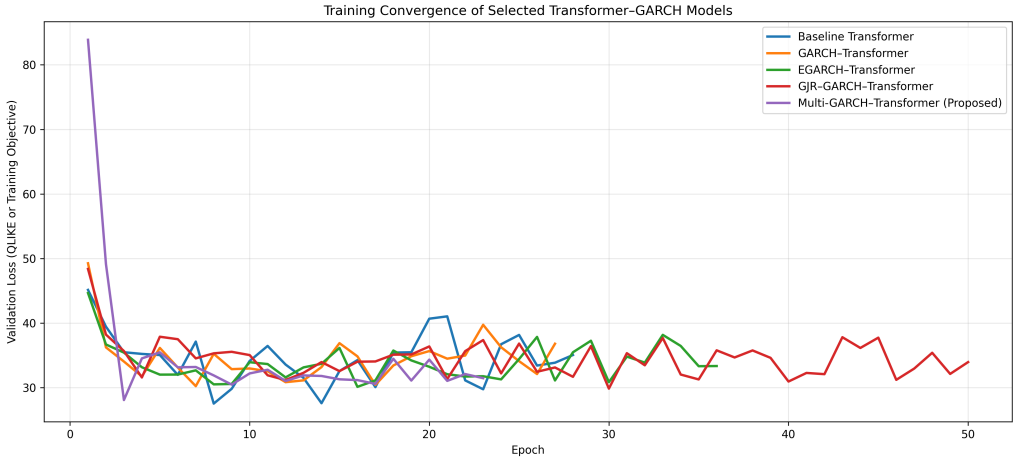

FIGURE 5. Validation loss trajectories during training.

Figure 6 shows that baseline Transformer models tend to under-react to abrupt volatility spikes, whereas the ensemble-informed hybrid model tracks realised volatility more closely and adapts more rapidly during regime transitions.

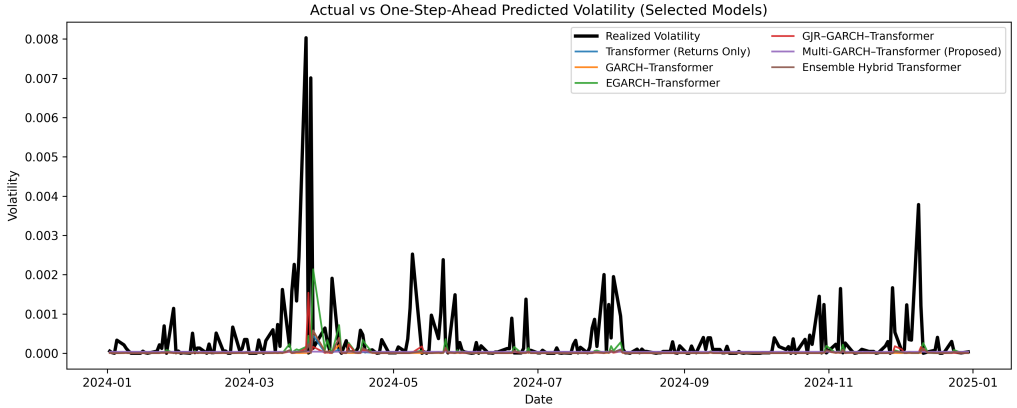

FIGURE 6. Actual versus predicted volatility for selected models.

3.4. Risk implications. Figure 7 links predictive volatility to economic interpretation through volatility-implied risk bands. Improved coverage during turbulent periods demonstrates enhanced robustness for risk-sensitive applications such as monitoring and hedging [5].

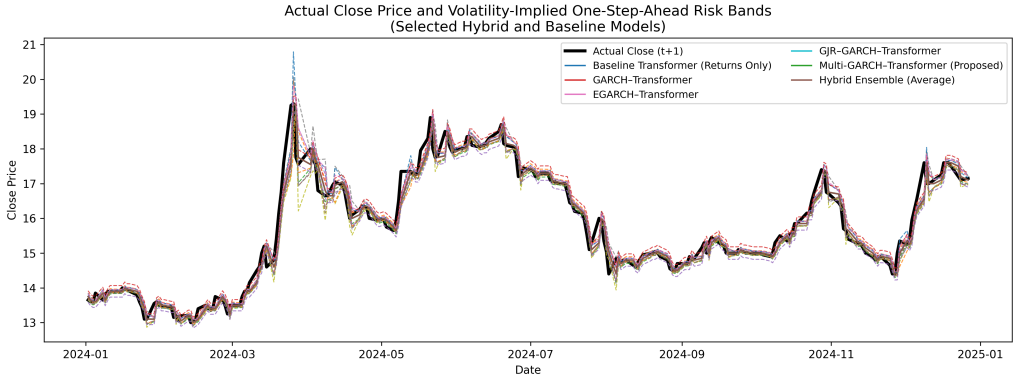

FIGURE 7. Volatility-implied risk bands.

3.5. Statistical comparison. Diebold-Mariano tests confirm that forecast improvements are statistically significant rather than arising from sampling variability [18].

3.6. Synthesis and contribution framing. The empirical results of this study demonstrate that embedding multiple econometric volatility signals directly into Transformer input representations enables representation-level ensemble learning that jointly models statistical structure and nonlinear temporal dynamics. By allowing attention mechanisms to adaptively weight heterogeneous volatility perspectives across time, the proposed modelling Architecture achieves improved forecasting stability, enhanced responsiveness to regime changes and statistically significant performance gains.

4. CONCLUSION, RECOMMENDATIONS, AND FUTURE DIRECTIONS

This study proposed a hybrid Multi-GARCH-Transformer architecture for volatility forecasting that integrates econometric modelling with Transformer-based deep learning through representation-level ensemble learning. By embedding multiple GARCH-family conditional variance estimates directly into the input representation, the proposed ensemble architecture combines structured econometric inductive biases with attention-based temporal representation learning. Unlike conventional hybrid approaches that aggregate model outputs post hoc, this

hybrid architecture performs ensemble integration within the representation space which enables the Transformer to adaptively weight heterogeneous volatility signals and jointly capture persistence, asymmetry and nonlinear temporal dependencies.

Results of this study demonstrate that the Multi-GARCH-Informed Transformer consistently achieves superior forecasting accuracy across volatility specific metrics, including QLIKE, RMSE and MAE, with statistically significant improvements confirmed through Diebold-Mariano testing. These gains arise from complementary inductive biases introduced by the ensemble of econometric models: persistence captured by GARCH, asymmetric magnitude sensitivity captured by EGARCH and leverage dynamics captured by GJR-GARCH. Integrating these heterogeneous signals within the attention mechanism allows the model to reconcile long-memory statistical structure with flexible nonlinear representation learning which improves stability and responsiveness to volatility regime shifts. The findings position representation-level ensemble learning as a principled approach for combining structured garch family models with modern attention-based architectures.

The findings of this study indicate that researchers and researchers should prioritise representation-level integration of complementary modelling architechtures rather than relying on single-model approaches or post hoc ensemble aggregation. Embedding multiple econometric volatility estimators as structured inputs allows deep learning architectures to leverage domain-specific inductive bias while retaining adaptive flexibility. For practical implementation, volatility sensitive evaluation architectures that emphasise QLIKE assessment, strict chronological validation and formal statistical comparison testing provide a more reliable foundation for performance evaluation and model selection.

This study examines only a Safaricom stock price return time series which limits direct inference regarding cross-asset or cross-market generalizability. Future research should evaluate hybrid multi-garch transformer architecture across diverse assets, market environments and volatility regimes to assess robustness under varying structural conditions. Incorporating exogenous information, including macroeconomic indicators, sentiment measures, or higher-frequency market features, may also improve predictive performance and economic interpretability. In addition, future research should extend the methodology toward probabilistic

forecasting and explicit uncertainty quantification to strengthen risk-sensitive decision-making. Further empirical validation across diverse settings is also needed to assess the broader potential of representation-level integration of econometric signals within attention-based models.

CONFLICT OF INTERESTS

The authors declare that there is no conflict of interests.

REFERENCES

[1] R.F. Engle, Autoregressive Conditional Heteroscedasticity with Estimates of the Variance of United Kingdom Inflation, Econometrica 50 (1982), 987–1007. https://doi.org/10.2307/1912773.

[2] A.J. Patton, Volatility Forecast Comparison Using Imperfect Volatility Proxies, J. Econ. 160 (2011), 246–256. https://doi.org/10.1016/j.jeconom.2010.03.034.

[3] Y. Bengio, A. Courville, P. Vincent, Representation Learning: A Review and New Perspectives, IEEE Trans. Pattern Anal. Mach. Intell. 35 (2013), 1798–1828. https://doi.org/10.1109/TPAMI.2013.50.

[4] O.B. Sezer, M.U. Gudelek, A.M. Ozbayoglu, Financial Time Series Forecasting with Deep Learning: A Systematic Literature Review: 2005–2019, Appl. Soft Comput. 90 (2020), 106181. https://doi.org/10.1016/j.asoc.2020.106181.

[5] C.T. Brownlees, R.F. Engle, Volatility, Correlation and Tails for Systemic Risk Measurement, SSRN (2011). https://doi.org/10.2139/ssrn.1611229.

[6] I. Goodfellow, Y. Bengio, A. Courville, Deep Learning, MIT Press, 2016.

[7] G. Zhang, Time Series Forecasting Using a Hybrid ARIMA and Neural Network Model, Neurocomputing 50 (2003), 159–175. https://doi.org/10.1016/S0925-2312(01)00702-0.

[8] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, et al., Attention Is All You Need, arXiv:1706.03762, 2017. https://doi.org/10.48550/arXiv.1706.03762.

[9] H. Zhou, S. Zhang, J. Peng, S. Zhang, J. Li, et al., Informer: Beyond Efficient Transformer for Long Sequence Time-Series Forecasting, Proc. AAAI Conf. Artif. Intell. 35 (2021), 11106–11115. https://doi.org/10.1609/aaai.v35i12.17325.

[10] J. Michańków, Ł. Kwiatkowski, J. Morajda, Combining Deep Learning and GARCH Models for Financial Volatility and Risk Forecasting, arXiv:2310.01063, 2023. https://doi.org/10.48550/arXiv.2310.01063.

[11] B. Lakshminarayanan, A. Pritzel, C. Blundell, Simple and Scalable Predictive Uncertainty Estimation Using Deep Ensembles, arXiv:1612.01474, 2016. https://doi.org/10.48550/arXiv.1612.01474.

[12] T. Bollerslev, Generalized Autoregressive Conditional Heteroskedasticity, J. Econ. 31 (1986), 307–327. https://doi.org/10.1016/0304-4076(86)90063-1.

[13] D.B. Nelson, Conditional Heteroskedasticity in Asset Returns: A New Approach, Econometrica 59 (1991), 347–370. https://doi.org/10.2307/2938260.

[14] L.R. Glosten, R. Jagannathan, D.E. Runkle, On the Relation between the Expected Value and the Volatility of the Nominal Excess Return on Stocks, J. Financ. 48 (1993), 1779–1801. https://doi.org/10.1111/j.1540-6261.1993.tb05128.x.

[15] S.H. Poon, C.W.J. Granger, Forecasting Volatility in Financial Markets: A Review, J. Econ. Lit. 41 (2003), 478–539. https://doi.org/10.1257/002205103765762743.

[16] W. Bao, J. Yue, Y. Rao, A Deep Learning Framework for Financial Time Series Using Stacked Autoencoders and Long-Short Term Memory, PLOS ONE 12 (2017), e0180944. https://doi.org/10.1371/journal.pone.0180944.

[17] T. Fischer, C. Krauss, Deep Learning with Long Short-Term Memory Networks for Financial Market Predictions, Eur. J. Oper. Res. 270 (2018), 654–669. https://doi.org/10.1016/j.ejor.2017.11.054.

[18] F.X. Diebold, R.S. Mariano, Comparing Predictive Accuracy, J. Bus. Econ. Stat. 13 (1995), 253–263. https://doi.org/10.1080/07350015.1995.10524599.