risks

Special Issue Reprint

Volatility Modeling in Financial Market

Edited by
Katarzyna Czech and Michał Wielechowski

mdpi.com/journal/risks

MDPI

Volatility Modeling in Financial Market

Volatility Modeling in Financial Market

Guest Editors

Katarzyna Czech
Michał Wielechowski

MDPI

Basel • Beijing • Wuhan • Barcelona • Belgrade • Novi Sad • Cluj • Manchester

Guest Editors

Katarzyna Czech
Department of Econometrics and Statistics
Warsaw University of Life Sciences-SGGW
Warsaw
Poland

Michał Wielechowski
Department of Economics and Economic Policy
Warsaw University of Life Sciences-SGGW
Warsaw
Poland

Editorial Office
MDPI AG
Grosspeteranlage 5
4052 Basel, Switzerland

This is a reprint of the Special Issue, published open access by the journal *Risks* (ISSN 2227-9091), freely accessible at: https://www.mdpi.com/journal/risks/special_issues/M0Q3OU19NL.

For citation purposes, cite each article independently as indicated on the article page online and as indicated below:

Lastname, Firstname, Firstname Lastname, and Firstname Lastname. Article Title. *Journal Name* **Year**, *Volume Number*, Page Range.

ISBN 978-3-7258-7993-9 (Hbk)
ISBN 978-3-7258-7994-6 (PDF)
https://doi.org/10.3390/books978-3-7258-7994-6

Contents

About the Editors

Katarzyna Czech

Katarzyna Czech, PhD, is an Assistant Professor at the Department of Econometrics and Statistics, Institute of Economics and Finance, Warsaw University of Life Sciences—SGGW. Her research interests focus on financial markets, volatility modelling, risk analysis, international market linkages, financial literacy, and the application of quantitative methods in economics and finance. In her research, she combines advanced econometric tools with the analysis of current economic and social challenges, including financial crises, the COVID-19 pandemic, geopolitical and political uncertainty, and asset price volatility. Dr Czech is the author and co-author of numerous scientific publications devoted to financial markets, exchange rates, market reactions to external shocks, volatility modelling, risk, and financial education. She has been actively involved in scientific and educational projects. She coordinated the Erasmus+ project "Learning by Experiencing Escape Rooms: Financial Literacy for Adults," financed by the European Commission. Dr Czech is also engaged in international academic cooperation, teaching, and editorial activities. She teaches courses in statistics, econometrics, dynamic and financial econometrics, financial mathematics, and financial engineering. As part of her teaching activities, she has also participated in several Erasmus+ mobility programs at various universities, including the University of British Columbia in Canada.

Michał Wielechowski

Michał Wielechowski, PhD, is an Assistant Professor in the Department of Economics and Economic Policy at the Institute of Economics and Finance, Warsaw University of Life Sciences. His research interests focus primarily on macroeconomics, economic policy, and financial markets. He is the author or co-author of more than 70 scientific publications and has presented numerous papers at conferences in Poland and abroad. He has participated in projects funded by the European Commission, the National Science Centre (NCN), the National Bank of Poland (NBP), and the Ministry of Education and Science. He has completed research fellowships at several institutions, including the National Bank of Poland (NBP), the University of Bonn, Wageningen University & Research, and SGH Warsaw School of Economics. He teaches courses in financial markets, investment portfolio, and economics. He is also a guest lecturer on the Global Economic Trends course at the Czech University of Life Sciences Prague. He has served as a lecturer under the Erasmus+ programme on multiple occasions in the Czech Republic, Greece, Hungary, Italy, and Spain. He serves as an expert for the Polish National Agency for Academic Exchange (NAWA) and the International Visegrad Fund.

risks

MDPI

Editorial

Special Issue "Volatility Modeling in Financial Market"

Katarzyna Czech [1,*] and Michał Wielechowski [2,*]

[1] Department of Econometrics and Statistics, Institute of Economics and Finance, Warsaw University of Life Sciences—SGGW, Nowoursynowska 166, 02-787 Warsaw, Poland
[2] Department of Economics and Economic Policy, Institute of Economics and Finance, Warsaw University of Life Sciences—SGGW, Nowoursynowska 166, 02-787 Warsaw, Poland
* Correspondence: katarzyna_czech@sggw.edu.pl (K.C.); michal_wielechowski@sggw.edu.pl (M.W.)

Dear Readers,

In this editorial, we provide a short overview of the Special Issue "Volatility Modelling in Financial Market" and summarise the main findings and contributions of the articles published within it. This Special Issue aims to bring together studies that advance the understanding of volatility, dependence, risk transmission, and forecasting in financial markets. We sincerely hope that this overview encourages readers to explore the full-length papers and engage further with the research questions addressed in this collection.

Volatility modelling remains one of the central areas of financial econometrics and risk management. Since volatility is not directly observable, researchers and practitioners rely on statistical, econometric, and computational methods to estimate and forecast it. In recent years, the field has evolved from classical ARCH- and GARCH-type models to more flexible approaches that incorporate realised measures, asymmetric dynamics, multivariate dependence structures, machine learning, soft computing, and hybrid frameworks. At the same time, the practical importance of volatility modelling has increased due to geopolitical shocks, changing market microstructure, the expansion of derivative markets, the rise of cryptocurrencies, and the growing use of artificial intelligence in financial decision-making.

Now that this Special Issue has been closed, we can state that it presents a broad and diverse set of contributions to volatility modelling and financial risk analysis. The published papers address different asset classes, including equities, exchange rates, corporate bonds, cryptocurrencies, derivatives, and credit portfolios. They also employ a wide range of methods, including GARCH-family models, DCC-GARCH, EGARCH, asymmetric GARCH specifications, realised-measure approaches, multiplicative error models, adaptive neuro-fuzzy inference systems, normalising flows, machine learning, and deep learning. Taken together, these studies show that volatility modelling is no longer limited to a single methodological tradition but is increasingly characterised by the integration of econometric interpretability with computational flexibility.

Several papers in this Special Issue contribute directly to the development and assessment of volatility forecasting methods. In "Historical Perspectives in Volatility Forecasting Methods with Machine Learning", Qiu et al. provide a comprehensive overview of the evolution of volatility forecasting, moving from implied volatility and GARCH-type models to recurrent neural networks, LSTMs, transformers, and other state-of-the-art learning-based approaches. Their results indicate that machine learning models can substantially improve forecasting accuracy, but they also underline important limitations related to interpretability, data requirements, computational cost, and robustness. This paper therefore offers a valuable bridge between classical financial econometrics and modern artificial intelligence methods.

Another methodological contribution is provided by Kirby in "Using Daily Stock Returns to Estimate the Unconditional and Conditional Variances of Lower-Frequency Stock Returns". This study shows that daily returns can be used to construct realised measures that are unbiased estimators of the unconditional and conditional variances of lower-frequency returns under relatively mild assumptions. The empirical analysis, based on S&P 500 data, suggests that multiplicative error models using these realised measures can outperform standard GARCH forecasts for weekly and monthly returns. This contribution is especially relevant when intraday data are unavailable or when long historical samples are required.

The paper "Normalising Flow Enhanced GARCH Models: A Two-Stage Framework for Flexible Innovation Modelling in Financial Time Series" by Hassan et al. proposes a hybrid NF-GARCH framework that preserves the interpretability of classical GARCH variance dynamics while replacing restrictive parametric innovation distributions with learned densities generated by normalising flows. The study shows that this approach can improve forecast accuracy, particularly for skewed-t GARCH baselines, while also providing more flexible modelling of heavy tails and asymmetric residual behaviour. At the same time, the authors point to computational costs and model-specific instability as important considerations for future research.

The Special Issue also includes papers focused on specific financial markets and instruments. Jongadsayakul's paper "Foreign Exchange Futures Trading and Spot Market Volatility in Thailand" examines whether the introduction of EUR/USD and USD/JPY futures on the Thailand Futures Exchange stabilises or destabilises the underlying spot market. Using EGARCH and VAR models, the study finds that the introduction of FX futures reduces spot volatility, increases the speed at which new information is incorporated into spot prices, and decreases the persistence of volatility shocks. The results also show that unexpected trading volume has a destabilising effect, while unexpected open interest has a stabilising effect, with the latter dominating overall.

In "Estimating Corporate Bond Market Volatility Using Asymmetric GARCH Models", Hadad et al. analyse the Israeli corporate bond market, which is characterised by high transparency and significant retail participation. The authors show that negative shocks have a stronger impact on volatility than positive shocks, confirming the importance of asymmetry and investor sentiment in corporate bond markets. Their findings indicate that the GJR-GARCH model with a Student's t-distribution best captures the volatility dynamics of the Tel-Bond 20 and Tel-Bond 60 indices. This contribution is particularly important because volatility behaviour in corporate bond markets remains less extensively studied than volatility in equity markets.

Czech and Wielechowski, in "Do Uncertainty and Action Shocks Affect G7 Stock Market Synchronisation? DCC-GARCH Evidence from the 2024 U.S. Election and the Reciprocal Tariffs Announcement", investigate how different types of U.S.-centred shocks affect conditional correlations between the U.S. equity market and other G7 markets. By distinguishing between an uncertainty shock represented by the 2024 U.S. presidential election and an action shock represented by the 2025 reciprocal tariff announcement, the study shows that different shocks produce different patterns of cross-market synchronisation. Election-related uncertainty is mainly associated with lower correlations for European markets, while the tariff-related action shock increases conditional correlations across all analysed U.S.–G7 pairs. This paper contributes to the literature on international spillovers, portfolio diversification, and state-dependent market co-movement.

The Special Issue also contains contributions that extend the discussion of volatility and risk modelling towards cryptocurrencies and credit risk. Orozco-Castañeda et al., in "Evaluating Volatility Using an ANFIS Model for Financial Time Series Prediction",

https://doi.org/10.3390/risks14050101

develop an ARIMA-ANFIS model for BTCUSD price prediction and risk assessment and compare it with an ARIMA-GARCH benchmark. Their findings suggest that ANFIS and GARCH capture different aspects of the data generation process. ANFIS may perform well under more stable conditions but can underestimate volatility during turbulent periods, whereas GARCH provides wider confidence intervals and stronger protection against high-volatility episodes. This study highlights the trade-off between flexibility and risk coverage in cryptocurrency volatility modelling.

Chang et al., in "Credit Risk Prediction Using Machine Learning and Deep Learning: A Study on Credit Card Customers", broaden the risk management perspective of the Special Issue by examining machine learning and deep learning techniques for credit default prediction. The study compares several models, including neural networks, logistic regression, AdaBoost, XGBoost, and LightGBM, and finds that XGBoost achieves the strongest predictive performance. Although this contribution does not focus on market volatility in the narrow sense, it complements the Special Issue by showing how machine learning tools can improve financial risk classification, lending decisions, and customer risk segmentation.

This Special Issue addresses several important gaps in the current literature and shows that classical GARCH-family models remain highly relevant, especially when interpretability, diagnostic testing, and practical risk management are central concerns. Moreover, it demonstrates that hybrid and machine learning approaches can improve predictive accuracy, but only when their limitations are carefully managed, including overfitting, computational complexity, data requirements, and reduced transparency. Additionally, the Special Issue highlights the need to model not only volatility levels but also asymmetry, tail behaviour, market synchronisation, innovation distributions, and the interaction between derivatives and spot markets. Finally, the collection confirms that volatility and risk dynamics are strongly context-dependent, varying across asset classes, market structures, investor compositions, and shock types.

In summary, the Special Issue "Volatility Modelling in Financial Market" presents a collection of studies that jointly contribute to the theoretical, methodological, and applied development of volatility and financial risk modelling. The published articles show that modern volatility research requires both rigorous econometric modelling and openness to new computational tools. By bringing together studies on GARCH models, realised measures, dynamic correlations, machine learning, soft computing, normalising flows, derivatives, corporate bonds, cryptocurrencies, equities, and credit risk, this Special Issue contributes to a deeper understanding of how financial risk is measured, transmitted, forecasted, and managed in increasingly complex markets.

Conflicts of Interest: The authors declare no conflicts of interest.

List of Contributions

Methodological and forecasting contributions

Provides a historical and comparative overview of volatility forecasting methods, from implied volatility and GARCH models to LSTM and transformer-based approaches:

Qiu, Zhiang, Clemens Kownatzki, Fabien Scalzo, and Eun Sang Cha. 2025. Historical Perspectives in Volatility Forecasting Methods with Machine Learning. *Risks* 13: 98. https://doi.org/10.3390/risks13050098.

Develops realised measures based on daily returns for estimating unconditional and conditional variances of lower-frequency stock returns:

Kirby, Chris. 2025. Using Daily Stock Returns to Estimate the Unconditional and Conditional Variances of Lower-Frequency Stock Returns. *Risks* 13: 190. https://doi.org/10.3390/risks13100190.

Introduces a two-stage NF-GARCH framework that combines classical GARCH variance dynamics with normalising-flow-based innovation modelling:

Hassan, Abdullah, Farai Mlambo, and Wilson Tsakane Mongwe. 2026. Normalising Flow Enhanced GARCH Models: A Two-Stage Framework for Flexible Innovation Modelling in Financial Time Series. *Risks* 14: 100. https://doi.org/10.3390/risks14050100.

Applied volatility and market-risk contributions

Examines the effect of FX futures trading on spot market volatility in Thailand using EGARCH and VAR models:

Jongadsayakul, Woradee. 2024. Foreign Exchange Futures Trading and Spot Market Volatility in Thailand. *Risks* 12: 107. https://doi.org/10.3390/risks12070107.

Analyses asymmetric volatility dynamics in the Israeli corporate bond market using GARCH-family models:

Hadad, Elroi, Amit Malka Fridman, and Rami Yosef. 2025. Estimating Corporate Bond Market Volatility Using Asymmetric GARCH Models. *Risks* 13: 224. https://doi.org/10.3390/risks13110224.

Investigates how uncertainty and action shocks affect G7 stock market synchronisation using DCC-GARCH models:

Czech, Katarzyna, and Michał Wielechowski. 2026. Do Uncertainty and Action Shocks Affect G7 Stock Market Synchronisation? DCC-GARCH Evidence from the 2024 U.S. Election and the Reciprocal Tariffs Announcement. *Risks* 14: 74. https://doi.org/10.3390/risks14040074.

Hybrid, soft-computing, and risk-prediction contributions

Applies an ARIMA-ANFIS framework to BTCUSD price prediction and volatility-related risk assessment:

Orozco-Castañeda, Johanna M., Sebastián Alzate-Vargas, and Danilo Bedoya-Valencia. 2024. Evaluating Volatility Using an ANFIS Model for Financial Time Series Prediction. *Risks* 12: 156. https://doi.org/10.3390/risks12100156.

Uses machine-learning and deep-learning methods for credit card default prediction and financial risk classification:

Chang, Victor, Sharuga Sivakulasingam, Hai Wang, Siu Tung Wong, Meghana Ashok Ganatra, and Jiabin Luo. 2024. Credit Risk Prediction Using Machine Learning and Deep Learning: A Study on Credit Card Customers. *Risks* 12: 174. https://doi.org/10.3390/risks12110174.

risks

MDPI

Article

Normalising Flow Enhanced GARCH Models: A Two-Stage Framework for Flexible Innovation Modelling in Financial Time Series

Abdullah Hassan [1], Farai Mlambo [2] and Wilson Tsakane Mongwe [1,*]

[1] School of Statistics and Actuarial Science, University of the Witwatersrand, Johannesburg 2050, South Africa; 1814643@students.wits.ac.za
[2] Graduate School of Business Administration, University of the Witwatersrand, Johannesburg 2050, South Africa; farai.mlambo@wits.ac.za
* Correspondence: wilsonmongwe@gmail.com

Abstract

We introduce the Normalising Flow GARCH (NF-GARCH), a two-stage hybrid framework that enhances traditional GARCH models by replacing restrictive parametric innovation distributions with learned densities via normalising flows. Our approach preserves the interpretability of standard variance dynamics while addressing the common issue of innovation misspecification. In the first stage, we estimate standard GARCH variants (sGARCH, TGARCH, and gjrGARCH) to extract standardised residuals. In the second stage, a Masked Autoregressive Flow learns the underlying residual distribution, with samples from the flow subsequently driving the GARCH recursion for out-of-sample forecasting. Evaluated on 13 daily financial series (six FX pairs and seven equities), NF-GARCH demonstrates systematic, statistically significant improvements in forecast accuracy for skewed-t baselines. Wilcoxon signed-rank tests confirm superior performance specifically for gjrGARCH-sstd and sGARCH-sstd specifications. While the framework offers enhanced flexibility and generative realism, we observe that computational overhead is increased, and the log-variance specification of eGARCH exhibits instability when paired with flow-based innovations. These results suggest that while NF-GARCH effectively captures empirical tail behaviour in univariate settings, future research should explore conditional flow architectures and multivariate extensions to account for time-varying innovation shapes. For risk management, gains are most relevant where skewed-t baselines are used and where closer residual realism supports scenario analysis; effect sizes remain modest relative to model risk and implementation cost.

Keywords: normalising flows; GARCH; volatility; financial time series; heavy tails; financial econometrics; stylized facts

1. Introduction

Empirical analysis over the past two decades has shown that financial returns exhibit clustering, where similarly large movements often follow large movements (Aloud et al. 2013; Cont 2001). These returns further exhibit heavy tails and asymmetric responses, with negative shocks increasing volatility more than positive shocks (Cont 2001). Generalised Autoregressive Conditional Heteroskedasticity (GARCH) models address clustering through autoregressive variance dynamics (Bollerslev 1986). However, their innovation

distributions are typically limited to Gaussian or Student's *t* forms. This limitation fails to capture the skewness and tail behaviour that are critical for risk assessment.

The choice of innovation distribution in GARCH models has long been recognised as critical to forecasting accuracy (Bollerslev 1987; Ederington and Guan 2005). Early specifications assumed Gaussian innovations, but this was quickly challenged by evidence of heavy tails (Bollerslev 1986). Bollerslev (1987) introduced Student's *t* innovations to capture excess kurtosis. Hansen (1994) demonstrated that even small innovation misspecifications propagate into substantial forecast errors. Despite advances with skewed distributions, parametric approaches impose restrictive functional forms (Fernández and Steel 1998).

Semi-parametric alternatives include kernel-based density estimation and nonparametric conditional density methods, though these suffer from curse-of-dimensionality issues and sensitivity to bandwidth choice (Engle and Ng 1993; Hansen 2004; Silverman 1986). Machine learning approaches have also been explored to enhance innovation modelling. Examples include Long Short-Term Memory networks integrated with GARCH, which can improve forecasting but sacrifice interpretability due to the black-box nature of neural network methods (Hossain and Nasser 2008; Kim and Won 2018). In regulated financial applications, interpretability and testable statistical properties remain important (Harvey et al. 2022; Rudin 2019).

Normalising flows provide a principled framework for learning complex probability distributions through a sequence of invertible transformations (Mongwe et al. 2025b; Rezende and Mohamed 2015). The base density is transformed via successive mappings using the change of variables formula, yielding a tractable yet expressive distribution. Normalising flows have the potential to address limitations of both parametric and alternative approaches to modelling innovations by providing exact likelihood evaluation through invertible transformations (Papamakarios et al. 2021; Rezende and Mohamed 2015; Seitz 2022). Despite extensive research, limited empirical evidence exists on whether flexible, nonparametric innovation distributions, specifically those learned via normalising flows, provide systematic improvements when integrated into classical GARCH frameworks without modifying the volatility recursion. This gap motivates a two-stage modular approach that we introduce in this paper.

Research gap and contributions. Semi-parametric and kernel-based GARCH specifications (Engle and Ng 1993; Hansen 2004) relax the innovation law but differ in implementation, bandwidth demands, and scalability. Joint (end-to-end) flow–GARCH estimation (Seitz 2022) entangles volatility and innovation learning, making it difficult to attribute forecast improvements to either component. We target the distinct question: if the volatility recursion is held fixed and estimated as in standard practice, does replacing the parametric residual law with a normalising flow improve forecasts and residual realism? Our specific contributions are: (i) a modular two-stage NF-GARCH design that preserves classical variance dynamics while learning a flexible innovation density; (ii) systematic evaluation on thirteen daily FX and equity series using out-of-sample metrics, Wilcoxon tests, distributional distances, VaR backtesting, and stress windows; (iii) empirical evidence that gains concentrate on skewed-*t* baselines (gjrGARCH-sstd, sGARCH-sstd) rather than Gaussian baselines, consistent with the hypothesis that flow-based density learning addresses residual skewness that parametric forms miss; (iv) demonstration that eGARCH instability under flow augmentation arises from parameter redundancy between the log-variance recursion and the flow's skewness modelling; and (v) a practical characterisation of when NF-GARCH should be preferred over standard GARCH in risk management contexts.

We propose Normalising Flow-GARCH (NF-GARCH), a hybrid approach that keeps the standard GARCH variance recursion but replaces parametric residuals with a nonparametric learned normalising flow. The learned innovation distribution better captures

https://doi.org/10.3390/risks14050100

empirical tail features while preserving interpretability of the variance dynamics. We evaluate four GARCH variants (sGARCH, TGARCH, GJR-GARCH, eGARCH) and their NF-augmented versions on 13 daily financial series (six FX pairs, seven equities) using chronological splits and time-series cross-validation. Metrics include out-of-sample loss (MSE, MAE), distributional distances, and VaR calibration.

The method operates in two stages. First, standard GARCH models are fitted and standardised residuals are extracted. Second, a normalising flow is trained on these residuals, and flow samples subsequently drive the original GARCH recursion. This preserves the traditional volatility structure and permits modular estimation. The flow's shape is state-invariant: it scales with σ_t but does not depend on it. End-to-end designs that jointly optimise flow and GARCH parameters represent a natural extension; we deliberately avoid joint estimation here to maintain interpretability and parameter identifiability, leaving this as a direction for future work.

The remainder of this paper is structured as follows. Section 2 presents the background to volatility modelling, normalising flows, and our two-stage hybrid approach. Section 3 describes the experiment setup. Section 4 presents the results and discussion, and Section 5 concludes.

2. Background

2.1. Volatility Modelling

The foundation of standard Autoregressive Moving Average (ARMA) analysis relies on the assumption that the mean and unconditional variance of the time series remain constant over time, implying stationarity (Box and Jenkins 1976; Pankratz 1991). Techniques such as plotting the Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF), alongside tests like the Augmented Dickey–Fuller (ADF) test, are employed to ascertain stationarity. For non-stationary time series, transformations are applied to achieve stationarity, as detailed by Pankratz (1991).

In addressing the challenge of high volatility in forecasting financial time series, recent developments have led to the adoption of models like Generalised Autoregressive Conditional Heteroskedasticity (GARCH) models. In GARCH processes, volatility (i.e., the variance of disturbances) is explicitly modelled. Tests for stationarity and other diagnostics closely parallel those used in ARMA analysis. Asymmetric extensions such as EGARCH (Nelson 1991), TGARCH (Zakoian 1994), and GJR-GARCH (Glosten et al. 1993) allow negative shocks to affect volatility differently from positive ones, capturing the leverage effect documented in equity and FX returns (Black 1976; Christie 1982; Cont 2001).

R. F. Engle (1982) showed that the serial correlation in squared returns, or conditional heteroskedasticity, can be modelled using an autoregressive conditional heteroskedasticity (ARCH) model of the following form:

$$Y_t = \mathbb{E}_{t-1}[Y_t] + \epsilon_t \tag{1}$$

$$\epsilon_t = \sigma_t z_t \tag{2}$$

$$\sigma_t^2 = a_0 + a_1\epsilon_{t-1}^2 + a_2\epsilon_{t-2}^2 + \cdots + a_q\epsilon_{t-q}^2 \tag{3}$$

where $\mathbb{E}_{t-1}[\cdot]$ represents the conditional expectation on all information that is available at time $t-1$ and ϵ_t is modelled as the product of a standardised shock z_t and time-varying volatility σ_t such that z_t is a sequence of independent and identically distributed ("iid") random variables with mean zero and unit variance. In the ARCH model, z_t is assumed to be independent and identically distributed with a standard normal distribution. The restrictions $a_0 > 0$ and $a_i >= 0$ $i = 1, \ldots, q$ are required for $\sigma_t^2 > 0$.

https://doi.org/10.3390/risks14050100

An important extension of the ARCH model proposed by Bollerslev (1986) replaces the AR (p) representation with an ARMA (p,q) formulation:

$$\sigma_t^2 = a_0 + \sum_{j=1}^{p} b_j \sigma_{t-j}^2 + \sum_{i=1}^{q} a_i \epsilon_{t-i}^2, \tag{4}$$

where the coefficients a_i $(i = 0, \ldots, q)$ and b_j $(j = 1, \ldots, p)$ are all assumed to be positive to ensure that the conditional variance σ_t^2 is always positive. The model in Section 2.4 together with Sections 2.1 and 2.2 is known as the generalised ARCH or GARCH (p,q) model. When $p = 0$, the GARCH model reduces to the ARCH model.

2.2. Normalising Flows

Normalising flows represent a robust framework for constructing intricate probability distributions by applying a series of invertible transformations (Mongwe et al. 2025a; Rezende and Mohamed 2015). The initial probability density undergoes a sequence of mappings by systematically employing the change of variables formula, yielding a tractable complex distribution. The term normalising flow aptly describes this process, as the density effectively "flows" through the sequence of transformations, culminating in a normalised probability distribution (Rezende and Mohamed 2015).

Based on the initial framework developed by Rezende and Mohamed (2015), we consider a finite sequence of transformations where each mapping is invertible and smooth. If we let f denote such a mapping with its corresponding inverse f^{-1}, then for a random variable $\mathbf{X}$ with an initial density $p_{\mathbf{X}}(\mathbf{x})$, the transformed random variable $\mathbf{Y} = f(\mathbf{X})$ follows a density

$$p_{\mathbf{Y}}(\mathbf{y}) = p_{\mathbf{X}}\left(f^{-1}(\mathbf{y})\right) \left| \det \frac{\partial f^{-1}(\mathbf{y})}{\partial \mathbf{y}} \right| \tag{5}$$

This relationship is derived by invoking the inverse function theorem, which governs the behaviour of Jacobians for invertible functions. By composing multiple such mappings, one can generate arbitrarily complex densities. Specifically, if a random variable $\mathbf{X}$ with density $p_{\mathbf{X}}(\mathbf{x})$ is subjected to a sequence of K transformations $f_1, f_2, \ldots, f_K$, the resulting density of the final variable $\mathbf{Z}_K = f_K \circ \cdots \circ f_1(\mathbf{X})$ can be written in terms of the base variable and the Jacobians of the forward maps $p_{\mathbf{Z}_K}(\mathbf{z}_K) = p_{\mathbf{X}}(\mathbf{x}_0) \prod_{k=1}^{K} \left| \det J_{f_k}(\mathbf{z}_{k-1}) \right|^{-1}$, where $\mathbf{z}_0 = \mathbf{x}_0$ and J_{f_k} denotes the Jacobian of f_k. Equivalently, in inverse form,

$$p_{\mathbf{Z}_K}(\mathbf{z}_K) = p_{\mathbf{X}}(\mathbf{x}) \prod_{k=1}^{K} \left| \det \frac{\partial f_k^{-1}(\mathbf{z}_k)}{\partial \mathbf{z}_k} \right| \tag{6}$$

As suggested by Kobyzev et al. (2020), the trajectory traced by the sequence of transformed random variables $\mathbf{Z}_k$, starting from the initial distribution $p_{\mathbf{X}}(\mathbf{x})$, constitutes the flow. In contrast, the sequence of intermediate densities $p_{\mathbf{Z}_k}(\mathbf{z}_k)$ defines the Normalising Flow.

2.3. Normalising Flows in GARCH Residuals (NF-GARCH)

Traditional GARCH-type models, including sGARCH, EGARCH, and TGARCH, often assume that the residuals z_t follow a known parametric distribution such as Gaussian, Student's t, or Generalised Error Distribution. However, empirical studies have shown that these fixed-distribution assumptions are often too restrictive to capture the true nature of financial return innovations, which can be skewed, heavy-tailed, or even multi-modal. The choice of innovation distribution is critical for tail-sensitive risk measures such as Value-at-Risk and Expected Shortfall (Bauwens et al. 2006; Hansen 1994).

https://doi.org/10.3390/risks14050100

To address this, we replace the assumption $z_t \sim \mathcal{N}(0,1)$ with the following:

$$z_t = f(u_t), \quad u_t \sim \mathcal{N}(0, I_d)$$

where $f : \mathbb{R}^d \to \mathbb{R}^d$ is a sequence of invertible, differentiable transformations—i.e., a normalising flow. This transforms a simple base distribution (such as standard normal) into a rich, learned distribution for residuals z_t, thereby enabling greater flexibility in capturing empirical data characteristics. While alternative heavy-tailed base distributions such as Student's t could be considered for the base distribution, we adopt a standard normal base for tractability and comparability, with tail flexibility introduced through the learned flow transformations.

Given a base density $p_U(u)$ and an invertible transformation $z = f(u)$, the transformed density of z is given by the change-of-variables formula:

$$p_Z(z) = p_U(f^{-1}(z)) \left| \det \left(\frac{\partial f^{-1}(z)}{\partial z} \right) \right|$$

or, equivalently,

$$\log p_Z(z) = \log p_U(u) - \sum_{i=1}^{K} \log \left| \det \left(\frac{\partial f_i}{\partial h_{i-1}} \right) \right|$$

where $f = f_K \circ \cdots \circ f_1$, and $h_i = f_i(h_{i-1})$ with $h_0 = u$. Popular choices for flows include RealNVP (Dinh et al. 2016), Masked Autoregressive Flows (MAF) (Papamakarios et al. 2017), and Neural Spline Flows (Durkan et al. 2019). Although these flows are commonly constructed using a standard normal base distribution, heavier-tailed behaviour can be accommodated through sufficiently expressive transformations, and alternative base distributions could be considered as extensions.

2.4. NF-GARCH Model Structure

The NF-GARCH model can be represented as follows:

$$Y_t = \mu_t + \epsilon_t \tag{7}$$

$$\epsilon_t = z_t \sigma_t, \quad z_t = f(u_t), \quad u_t \sim \mathcal{N}(0,1) \tag{8}$$

$$\sigma_t^2 = \omega + \sum_{i=1}^{q} \alpha_i \epsilon_{t-i}^2 + \sum_{j=1}^{p} \beta_j \sigma_{t-j}^2 \tag{9}$$

Empirical specification (notation). The system above uses general orders (p, q) for exposition. In all experiments we estimate constant-mean log-return models, $r_t = \mu + \epsilon_t$, where the scalar μ is estimated by maximum likelihood jointly with volatility parameters (no ARMA dynamics in the mean). The conditional variance follows each variant's **GARCH(1,1)**-type recursion as implemented in our custom R engine: sGARCH updates σ_t^2 via the standard GARCH(1,1) recursion; TGARCH models the conditional scale σ_t directly per Zakoian (1994); GJR-GARCH adds an asymmetric leverage term $I_{t-1}\epsilon_{t-1}^2$ to the variance recursion; and eGARCH models $\log \sigma_t^2$ with asymmetric terms per Nelson (1991). Standardised residuals passed to the flow are $\hat{z}_t = (r_t - \hat{\mu})/\hat{\sigma}_t$ using in-sample fitted values only. In tables, the label **norm** denotes Gaussian innovations; **sstd** denotes skewed Student's t innovations in the parameterisation of Fernández and Steel (1998) and Hansen (1994).

Figure 1 highlights the modular nature of the NF-GARCH framework: the normalising flow modifies only the innovation distribution, while the volatility recursion remains structurally identical to classical GARCH models.

https://doi.org/10.3390/risks14050100

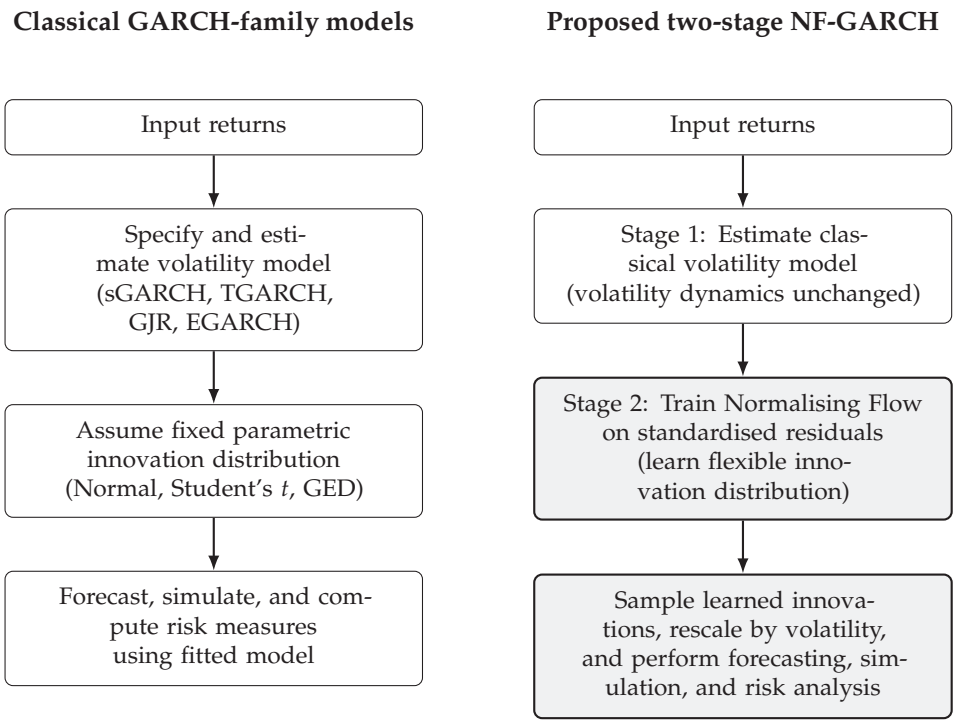


Figure 1. Comparison of classical GARCH-family workflows and the proposed two-stage Normalising Flow–GARCH framework.

In principle, the parameters of both the volatility recursion and the flow can be estimated by maximising the likelihood implied by the GARCH recursion combined with the flow-induced residual density $p_Z(z)$. This framework allows the volatility process to follow the standard GARCH formulation while modelling the innovation distribution as a flexible, non-Gaussian transformation of a simple base distribution.

2.5. Benefits of NF-GARCH

The integration of normalising flows into GARCH models enables the capture of skewness and heavy tails through flexible invertible transformations, without imposing restrictive parametric constraints on the innovation law, while preserving exact likelihoods for estimation. The resulting residual distribution facilitates realistic simulation and stress testing. Notably, the volatility process remains interpretable because the variance recursion is unchanged, even as the flexibility of the residuals increases. This hybrid framework isolates the contribution of distributional flexibility without altering the volatility structure, thereby allowing the assessment of whether a more expressive innovation law enhances forecasting accuracy, scenario generation, and the replication of stylized facts.

This hybrid approach allows us to assess whether distributional flexibility alone (i.e., with no change to the volatility structure) can significantly improve forecasting, scenario simulation, and stylized fact replication.

2.6. Theoretical Considerations

Normalising flows offer a flexible and tractable approach for modelling innovation distributions; however, their integration with GARCH-family volatility recursions introduces several important theoretical considerations.

2.6.1. Identifiability

Within the NF-GARCH framework, the conditional variance σ_t^2 is determined by the GARCH recursion, while the flow $f_\theta(\cdot)$ transforms latent noise u_t into residuals z_t.

https://doi.org/10.3390/risks14050100

There is a potential for a confounding effect between the volatility parameters and the transformations learned by the flow as both components influence the scale and shape of the simulated return given that

$$r_t = \mu_t + \epsilon_t, \qquad \epsilon_t = \sigma_t z_t, \qquad z_t = f_\theta(u_t),$$

If the flow is excessively flexible, it may absorb variation that would otherwise be attributed to the volatility recursion, leading to partial non-identifiability between scale parameters in σ_t and in f_θ. Formally, writing $r_t = \sigma_t f_\theta(u_t)$, if f_θ includes an arbitrary scaling component then $\sigma_t f_\theta(u_t) = (\sigma_t c)\tilde{f}_\theta(u_t)$ for some constant c, so scale can be reallocated between the volatility recursion and the flow, implying partial scale non-identifiability. This is an inherent limitation of combining highly expressive innovation laws with parametric volatility models, which becomes even more important when one wants to jointly calibrate the parameters of the underlying GARCH model with that of the Normalising Flow.

2.6.2. Overfitting in Expressive Flows

As with any deep learning model, normalising flows with numerous layers or high-capacity coupling networks are susceptible to overfitting the empirical residual distribution, particularly when applied to short financial time series (Mongwe et al. 2025b). Since flows are trained by maximising likelihood, they may capture spurious high-frequency structure or noise instead of the underlying innovation law. In volatility modelling, such overfitting can distort tail behaviour, degrade out-of-sample forecast performance, and yield misleading risk measures. Therefore, controlling model capacity, such as by limiting flow depth or width and employing regularisation techniques, is essential in NF-GARCH design (Mongwe et al. 2025b).

2.6.3. Likelihood Regularisation and Numerical Stability

Flow-based likelihoods require computation of the Jacobian log-determinant,

$$\log p_Z(z) = \log p_U\left(f_\theta^{-1}(z)\right) + \log\left|\det J_{f_\theta^{-1}}(z)\right|,$$

which can become numerically unstable if the transformations are too deep or poorly conditioned. Large positive or negative Jacobian terms can lead to exploding or vanishing log-densities and hinder convergence in optimisation (Papamakarios et al. 2021). Practical implementations often mitigate these issues through weight decay, spectral constraints, or penalties on extreme Jacobian values to keep the learned density well-behaved (Papamakarios et al. 2021).

2.6.4. Sensitivity to Flow Depth and Architecture

While deeper flows are theoretically more expressive, Papamakarios et al. (2021) confirmed that they also increase computational cost and may interact unfavourably with long-memory volatility dynamics. In practice, shallow or moderately deep flows often provide sufficient flexibility to capture skewness and heavy tails in the residuals, whereas very deep architectures yield diminishing returns and intensify identifiability and stability concerns (Liu and Regier 2020). Consequently, parsimonious flow architectures are generally preferable when integrating flows into GARCH-type models.

2.6.5. Implications for NF-GARCH

These considerations indicate that NF-GARCH models function as semi-parametric volatility models, with their advantages contingent upon achieving a careful balance between flexibility and control. The GARCH recursion maintains its interpretability, whereas

https://doi.org/10.3390/risks14050100

the innovation law becomes a learned component that requires regularisation to prevent overfitting and instability. Thus, a theoretically robust NF-GARCH specification necessitates careful attention to both the volatility structure and the capacity and regularisation of the flow.

2.7. Two-Stage NF Innovations vs. End-to-End Flow-After-Scaling

Two-stage. In this framework, the normalising flow is trained after the standard GARCH estimation stage, rather than being embedded directly into the volatility recursion. This modular approach separates the estimation of conditional volatility dynamics from the learning of the innovation distribution, allowing for a clearer interpretation of each component. Specifically, we model the following:

$$r_t = \mu_t + \epsilon_t, \qquad \epsilon_t = \sigma_t z'_t, \quad z'_t \sim p_{\mathrm{NF}},$$

with σ_t^2 following the usual GARCH recursion and p_{NF} learned from the standardised residuals $\hat{z}_t = (r_t - \hat{\mu}_t)/\hat{\sigma}_t$. Estimation is modular: first fit a GARCH model under a standard innovation assumption (Gaussian or skew-t) to obtain $\hat{\sigma}_t$, then train a flow on $\{\hat{z}_t\}$. Forecasts and simulations draw $z'_t \sim p_{\mathrm{NF}}$ and propagate through the same recursion.

Joint (end-to-end) training. An alternative modelling strategy would estimate the GARCH volatility parameters and the normalising flow jointly within a single likelihood function, which is a flow-after-scaling or end-to-end calibration approach. In such a framework, the innovations are modelled as

$$z_t = f_\phi(u_t), \quad u_t \sim \mathcal{N}(0,1), \qquad \epsilon_t = \sigma_t(\theta) z_t,$$

where both (θ, ϕ) are estimated simultaneously by maximising the joint log-likelihood $\sum_t \ell_t(\theta, \phi)$. This allows the flow to adapt to the evolving volatility state and enables fully data-driven conditional innovation laws, but it introduces a highly non-convex objective and greater computational cost.

While theoretically appealing, this approach introduces substantial practical and conceptual challenges. Joint calibration entangles the scale effects of the volatility recursion with the shape flexibility of the flow, leading to partial identifiability between $\sigma_t(\theta)$ and $f_\phi(\cdot)$. As a result, improvements in likelihood or forecasting accuracy cannot be uniquely attributed to enhanced volatility dynamics or improved innovation modelling. In addition, the resulting optimisation problem is highly non-convex and numerically unstable, particularly for asymmetric volatility specifications such as EGARCH.

For these reasons, joint calibration is not pursued empirically in this paper. Joint estimation was not pursued due to numerical instability and the need to isolate the marginal contribution of innovation modelling. Instead, a two-stage design is adopted to deliberately isolate the contribution of flexible innovation distributions while preserving the classical GARCH volatility structure. Nevertheless, joint flow–GARCH estimation remains an important and promising direction for future research, particularly in settings where conditional density calibration is prioritised over interpretability.

2.8. Relationship to Classical GARCH

The proposed NF-GARCH framework strictly generalises the classical GARCH model. In particular, standard GARCH with Gaussian innovations is recovered as a special case when the normalising flow reduces to the identity transformation

$$f(u) = u,$$

https://doi.org/10.3390/risks14050100

so that

$$z_t = f(u_t) = u_t \sim \mathcal{N}(0,1).$$

Under this restriction, the NF-GARCH model collapses exactly to the conventional GARCH specification with Gaussian residuals. Hence, classical GARCH models lie on the boundary of the NF-GARCH model class.

It is important to note, however, that this nesting is conceptual rather than operational: in the two-stage framework adopted in this paper, the flow is not jointly estimated with the volatility recursion, and the identity mapping is not imposed during training. As such, standard GARCH is not recovered through estimation but represents a limiting case within the broader NF-GARCH modelling space. A comparison of innovation structures in standard GARCH and in the NF-GARCH framework are shown in Figure 2.

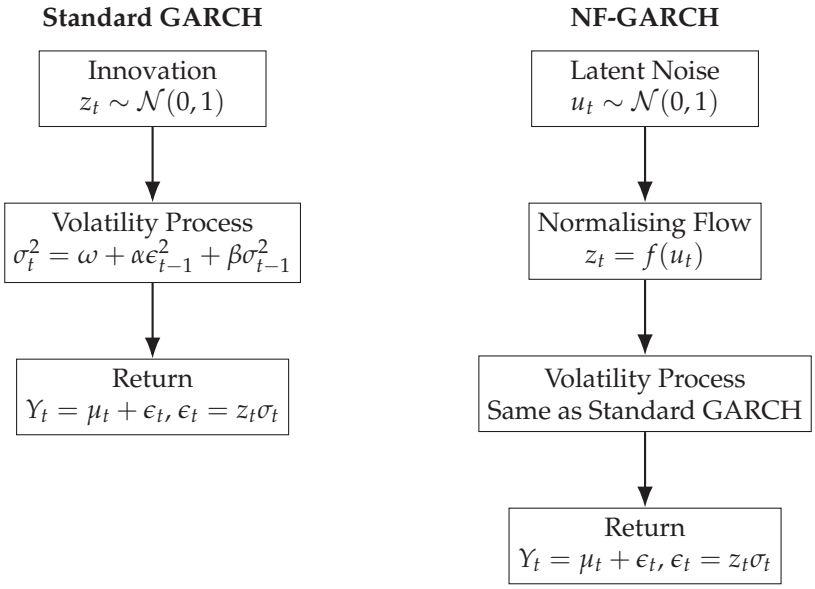


Figure 2. Comparison of innovation structures in Standard GARCH vs. NF-GARCH frameworks.

3. Experiment Setup

3.1. Data Description

Our empirical analysis utilises a dataset comprising 13 daily financial time series, spanning the period from 31 August 2005, to 31 August 2024. The sample encompasses six foreign exchange (FX) pairs (USD/ZAR, GBP/USD, EUR/USD, GBP/CNY, GBP/ZAR, and EUR/ZAR) alongside seven highly capitalized equities (X, NVDA, MSFT, PG, CAT, WMT, and AMZN). FX data were obtained from the South African Reserve Bank and OANDA, while equity price data were retrieved from Yahoo Finance. All series were systematically aligned to exclude non-trading days. Because this sample predominantly represents highly liquid, major asset classes and widely traded emerging market currencies, we caution that the empirical findings may not directly generalise to alternative asset classes such as commodities, fixed-income instruments, or illiquid frontier markets.

Asset selection rationale. Assets were selected to span two liquid classes—major and ZAR-denominated FX pairs and large-cap multi-sector equities—that are well known to exhibit GARCH-type conditional heteroskedasticity and heavy-tailed innovations. Including ZAR pairs (USD/ZAR, GBP/ZAR, EUR/ZAR) introduces emerging-market currency dynamics alongside G10 benchmarks (GBP/USD, EUR/USD). The equities cover technology (NVDA, MSFT), consumer staples (PG, WMT), industrials (CAT, X), and e-commerce (AMZN), providing sector diversity. This composition allows comparison across asset

https://doi.org/10.3390/risks14050100

classes while remaining computationally feasible; extension to commodities, fixed income, and frontier markets is noted as a direction for future work.

Returns and cleaning. We work with log returns $r_t = \log(P_t/P_{t-1})$ computed on business days. Prices are aligned to each asset's trading calendar; non-trading days and missing values are dropped. Extreme single-day spikes are retained without winsorisation so that tail behaviour is preserved for flow training and VaR analysis. A minimum of 520 valid observations is required per asset (500 training + 20 test); assets or windows that fail to reach this threshold are excluded. Return series are treated as weakly stationary following standard practice, with formal residual diagnostics after GARCH filtering reported in Section 3.5.

3.2. Data Preparation and Performance Evaluation

To preserve the inherent temporal dynamics of the sample, we partition the dataset chronologically, allocating 65% to the training set and 35% to the test set. Robustness is further established through a rolling-window time-series cross-validation framework. Within this framework, each cross-validation fold comprises a training window of 500 observations followed immediately by a 20-observation test window, advancing in discrete, non-overlapping 500-observation increments. To manage computational feasibility while ensuring representative temporal coverage, we subsampled up to three evenly spaced folds per asset–model pair. Consequently, these validation windows strategically capture the early, mid, and late evaluation periods across all 13 analysed assets and their respective models.

To strictly preclude look-ahead bias, we enforce strict chronological ordering: each training window spans the interval from t_{start} to $t_{\text{start}} + 499$, with the subsequent 20 observations designated for out-of-sample testing. Consequently, a minimum threshold of 520 valid observations is required for an asset to be included in the sample. Furthermore, the final analysis is restricted exclusively to models that successfully achieve convergence. To evaluate predictive accuracy and model fit, we calculate the mean squared error (MSE), Mean Absolute Error (MAE), Akaike Information Criterion (AIC), and Bayesian Information Criterion (BIC) independently for each window before computing their overall averages. This localized evaluation framework captures performance heterogeneity across diverse market regimes and mitigates reliance on any single temporal period.

3.3. Model Development and Implementation

The GARCH models are implemented in R (version 4.4.1; R Core Team, Vienna, Austria) using a custom estimation engine. This gives complete control over parameterisation and diagnostics. We fit four variants per asset: standard GARCH (normal and Student-t innovations), Exponential GARCH (asymmetric effects), Glosten–Jagannathan–Runkle GARCH (leverage terms), and Threshold GARCH (regime-dependent responses). Maximum likelihood estimation uses R's `optim` function. We evaluate models using AIC, BIC, MSE, MAE, and residual diagnostics (Ljung–Box and ARCH–LM).

For the normalising flow, we extract standardised residuals from each fitted GARCH model and train an independent flow model using the `nflows` (version 0.14.0) library in Python (version 3.14). The main experiments use a Masked Autoregressive Flow (MAF); an architecturally distinct Real NVP model is trained under identical hyperparameters for the multi-seed robustness comparison. Architecture details and a side-by-side specification appear in Table 1.

https://doi.org/10.3390/risks14050100

Table 1. Normalising flow architecture specifications †.

Component	MAF (Main)	RealNVP (Robustness)
Architecture		
Transform Type	Masked Autoregressive Flow	Real Non-Volume Preserving
Coupling	Autoregressive affine; masked connections	Non-autoregressive affine coupling
Masking	Autoregressive ordering	Input split into two halves
Residual Blocks	—	2 per coupling layer
Shared hyperparameters		
Base Distribution	Standard Normal $\mathcal{N}(0,1)$	
Layers	4	
Hidden Units	64 per layer	
Activation	ReLU	
Optimizer	Adam	
Learning Rate	10^{-3}	
Batch Size	512	
Max. Epochs	75	
Early Stopping	Patience = 15 epochs (validation log-likelihood)	
Validation Split	20% of training data	
Weight Decay	10^{-5}	

† *Note:* Both architectures use the `nflows` Python library (MAF: `MaskedAffineAutoregressiveTransform`; RealNVP: `RealNVP` class). MAF is used for all main experiments; RealNVP is applied with identical shared hyperparameters for the multi-seed robustness comparison (Section 4.9). RealNVP replaces autoregressive masking with non-autoregressive affine coupling layers, preserving exact invertibility and tractable log-determinant computation. —: not applicable (residual blocks apply to RealNVP coupling layers only; MAF does not use them).

Rationale for MAF. Among flow architectures, Masked Autoregressive Flows (Papamakarios et al. 2017) are a standard choice for low-dimensional univariate density estimation: they yield analytically tractable Jacobians, numerically stable maximum-likelihood training on residuals, and an autoregressive factorisation that naturally matches the univariate innovation setting. Real NVP (Dinh et al. 2016) and Neural Spline Flows (Durkan et al. 2019) are architecturally distinct alternatives mentioned in the paper; we adopt MAF for the main experiments after a constrained sensitivity check over depth, width, learning rate, and batch size (Section 3.4), prioritising validation likelihood and numerical stability. We extend the empirical comparison to Real NVP (Dinh et al. 2016) on the same evaluation protocol and report multi-seed stability results across seeds 123, 456, and 789 for both architectures in Section 4.9.

MAF transforms the base distribution through a series of invertible autoregressive layers (Papamakarios et al. 2017). Each layer applies an affine transformation $y_i = x_i \cdot \exp(s_i(\mathbf{x}_{<i})) + t_i(\mathbf{x}_{<i})$, where the scale s_i and shift t_i are conditioned only on preceding dimensions, ensuring strict invertibility and analytically tractable Jacobians. Post-training, NF-GARCH models are synthesised for simulation and forecasting by integrating the flow-generated innovations into the original GARCH recursion. While this study is confined to a univariate framework, extending the methodology to capture joint market dynamics via multivariate models (e.g., Multivariate GARCH (Bollerslev 1990), BEKK-MGARCH (Engle and Kroner 1995), or Dynamic Conditional Correlation (R. Engle 2002)) remains a compelling direction for future work.

Note that the described two-stage procedure prevents data leakage as follows:

1. **Data splitting:** Split chronologically—65% training, 35% test.
2. **Stage 1—GARCH estimation:**

https://doi.org/10.3390/risks14050100

- Estimate GARCH parameters on training set only: $\boldsymbol{\theta}^*_{\text{GARCH}} = \arg\max_{\boldsymbol{\theta}} \sum_{t \in \text{train}} \ell_{\text{GARCH}}(r_t; \boldsymbol{\theta})$;
- Extract standardised residuals: $\hat{z}_t = (r_t - \hat{\mu}_t)/\hat{\sigma}_t$ for $t \in$ train, where $\hat{\sigma}_t$ uses $\boldsymbol{\theta}^*_{\text{GARCH}}$.

3. **Stage 2—NF training:**
   - Train flow on training-set residuals only: $\boldsymbol{\phi}^* = \arg\max_{\boldsymbol{\phi}} \sum_{t \in \text{train}} \log p_f(\hat{z}_t; \boldsymbol{\phi})$;
   - Test-set information never enters NF training.
4. **Stage 3—NF-GARCH simulation and forecasting:**
   - Sample residuals from trained NF: $\tilde{z}_t \sim p_f(\cdot; \boldsymbol{\phi}^*)$.;
   - Forecast on test set: $\tilde{r}_t = \hat{\mu}_t + \tilde{z}_t \hat{\sigma}_t$ for $t \in$ test;
   - Evaluate using test-set returns only.

Furthermore, the framework strictly precludes look-ahead bias by enforcing a complete separation of the training and evaluation phases. Specifically, the GARCH parameters and their corresponding residuals are estimated exclusively on the training sample. The normalising flow is subsequently trained solely on these in-sample residuals, reserving the test set strictly for out-of-sample evaluation. This sequential methodology inherently assumes residual stationarity across both the training and test periods—an assumption we rigorously validate using Augmented Dickey–Fuller (ADF), KPSS, Ljung-Box, and ARCH-LM tests (see Section 3.5).

3.4. Hyperparameter Selection and Model Capacity Control

The implementation of normalising flows necessitates the specification of various architectural and optimisation hyperparameters. We followed a two-phase selection protocol. First, the final 20% of each training window was withheld as a held-out validation set before any hyperparameter search began, ensuring no overlap with the test period. Second, we evaluated candidate configurations over network depth (3–6 layers), width (32–128 hidden units), learning rates ($[5 \times 10^{-4}, 2 \times 10^{-3}]$), and batch sizes (256–1024); the configuration achieving the highest mean validation log-likelihood across assets and GARCH specifications was retained. The finalised Masked Autoregressive Flow (MAF) architecture comprises four layers with 64 hidden units each, using a batch size of 512. The network is optimised via the Adam algorithm with a learning rate of 10^{-3}, incorporating an early stopping mechanism triggered after 15 epochs without improvement in validation log-likelihood. The same hyperparameters were applied without re-tuning to the RealNVP robustness comparison in Section 4.9, providing a controlled architectural comparison.

Notably, expanding the network's capacity—either in depth or width—yielded negligible in-sample improvements while inducing erratic fluctuations in the Jacobian log-determinants and pronounced overfitting, particularly on shorter time series. To mitigate these instabilities, we explicitly restricted the architectural depth and applied L_2 regularisation with a weight decay parameter of 10^{-5}.

3.5. Residual Stationarity Diagnostics

The two-stage approach assumes standardised residuals

$$\hat{z}_t = \frac{r_t - \hat{\mu}_t}{\hat{\sigma}_t}$$

are weakly stationary after GARCH filtering, suitable for density estimation. We test this for each asset–model pair using: Augmented Dickey–Fuller (unit root), KPSS (stationarity), autocorrelation functions for $\hat{z}_t$ and $\hat{z}_t^2$, Ljung–Box Q-statistics for raw and squared residuals, and ARCH-LM tests for remaining heteroskedasticity.

https://doi.org/10.3390/risks14050100

ADF tests reject unit roots for all models and assets. KPSS indicates stationarity for most series. GARCH filtering largely removes serial dependence in $\hat{z}_t$; some squared residuals show minor remaining dependence. ARCH-LM shows substantial—but not complete—heteroskedasticity reduction.

Table 2 summarises the diagnostics. Most series pass stationarity tests (ADF, KPSS) with mean p-values strongly against non-stationarity. Ljung-Box shows most residuals are approximately uncorrelated; some squared residuals retain dependence. ARCH-LM indicates substantially reduced heteroskedasticity, though some series keep mild ARCH effects. These support the weak stationarity assumption for flow-based density estimation, while acknowledging that complete heteroskedasticity removal is not universal.

Table 2. Summary of residual stationarity diagnostics.

Test	Mean p-Value	Pass Rate (%)	Interpretation
ADF (unit root)	0.012	94.4	Most residuals reject non-stationarity
KPSS (stationarity)	0.087	88.9	Most residuals fail to reject stationarity
Ljung-Box ($\hat{z}_t$)	0.156	83.3	Most residuals show no serial correlation
Ljung-Box ($\hat{z}_t^2$)	0.089	77.8	Some squared residuals retain dependence
ARCH-LM	0.124	72.2	Most residuals show reduced heteroskedasticity

Note: Pass rates indicate the percentage of model–asset combinations where tests support the null hypothesis (for ADF, KPSS, Ljung-Box) or fail to reject homoskedasticity (for ARCH-LM). Lower p-values for ADF and KPSS, and higher p-values for Ljung-Box and ARCH-LM, indicate better residual properties. Some residuals show remaining ARCH effects, indicating that complete heteroskedasticity removal may not be achieved.

3.6. Theoretical Properties of Two-Stage Estimation

Two-stage estimation needs theoretical justification. Under regularity conditions (Newey and McFadden 1994), consistency requires: (1) Stage 1 (GARCH) is consistent, (2) Stage 2 (NF) depends on Stage 1 only through residuals, and (3) residual distribution is stationary across training and test periods.

Two-stage is less efficient than joint maximum likelihood as it ignores cross-equation information between GARCH and flow parameters. Practical advantages: separate optimisation (computationally tractable), numerical stability (avoids high-dimensional joint optimisation), and flexibility (different methods per stage). Efficiency loss is typically small when residuals are approximately stationary, which we validate via diagnostics; formal efficiency comparison with joint one-step extremum estimators is left for future work.

Validity requires: (1) GARCH and NF parameters are separately identifiable, (2) residual distribution is stationary (ADF/KPSS), and (3) no functional dependence between GARCH and NF parameter spaces. Identifiability holds because GARCH controls volatility dynamics while NF controls residual shape, which are distinct model aspects.

3.7. Conditional vs. Unconditional Innovation Modelling

Our two-stage design models the standardised residuals using a time-invariant unconditional distribution, acknowledging that financial innovations may occasionally exhibit mild conditional heterogeneity during periods of acute market stress. In this framework, all temporal dynamics in volatility are captured exclusively by the GARCH recursion via the conditional standard deviation, σ_t. Concurrently, the normalising flow specifies a highly flexible, yet time-invariant, distribution for the innovations. This architectural choice structurally mirrors classical GARCH specifications, wherein innovations are assumed to be independent and identically distributed (i.i.d.) according to a fixed parametric law (e.g., Gaussian, Student's t, or Generalised Error Distribution). The NF-GARCH model

https://doi.org/10.3390/risks14050100

retains this foundational structure but substitutes the rigid parametric density with a data-driven, highly parameterised distribution learned via the flow.

To validate the assumption of a time-invariant innovation distribution, we conducted a battery of robustness checks, including rolling-variance metrics, structural break tests, and subsample distributional comparisons. The results confirm that while the underlying return series exhibit the expected time-varying conditional variance, the standardised residual distributions display no substantial structural breaks and demonstrate limited instability, thereby empirically supporting our time-invariant modelling approach. Although fully conditional or state-dependent normalising flows could theoretically capture more complex distributional dynamics, such extensions introduce significant identifiability and optimisation challenges, rendering them a subject for future research.

3.8. Model Evaluation and Predictions

To rigorously assess the out-of-sample predictive accuracy and distributional fidelity of the final models, we implement a comprehensive dual-evaluation framework on a held-out test set. The first phase evaluates the quality of the synthetic residuals generated by the normalising flows. We employ a robust suite of distributional diagnostics, comprising the (1) Kolmogorov–Smirnov statistic, (2) Wasserstein-1 distance, (3) the Hill estimator for tail indices, (4) skewness, and (5) kurtosis to verify that the flow-injected innovations successfully replicate the key empirical properties of the underlying financial data.

The second phase quantifies explicit forecasting performance utilising the previously outlined 65/35 chronological split and rolling-window time-series cross-validation. Model efficacy is measured via the mean squared error (MSE), Mean Absolute Error (MAE), out-of-sample log-likelihood, AIC, and BIC. To determine the statistical significance of performance differentials across matched model–asset–window evaluations, we apply the non-parametric Wilcoxon signed-rank test (Demšar 2006). Furthermore, forecast win-rates are computed to provide an interpretable summary of relative model superiority, aligning with established best practices in volatility forecasting (Amisano and Giacomini 2007; Ziel 2016). Finally, through a combination of visual diagnostics and these formal distributional metrics, we systematically verify the replication of stylized financial facts. This confirms that the performance gains observed in the NF-augmented models represent robust, structural improvements rather than transient initialisation artifacts.

3.9. Stress Testing Framework and Scenario Analysis

To rigorously evaluate the predictive stability of the proposed framework under extreme market conditions, we implement a comprehensive stress-testing protocol encompassing historical crisis scenarios. The historical analysis examines two periods of acute market dislocation: the Global Financial Crisis (1 September 2008–31 March 2009) and the COVID-19 market crash (1 February 2020–30 April 2020). To maintain strict out-of-sample integrity, models are calibrated exclusively on pre-crisis data and subsequently evaluated during these defined crisis windows. Forecasting performance across all scenarios is quantified via MSE and MAE. We acknowledge that the scope of this evaluation is limited to two specific historical crises. While incorporating supplementary historical stress periods and formally integrating regime-switching tests would further validate the model's robustness, such computationally intensive extensions remain vital avenues for future research.

4. Results and Discussion

In this section, we present the results of the experiments. Figure 3 shows the autocorrelation and partial autocorrelation functions of the EURUSD (FX) and NVDA (equity) instruments respectively as examples. Figures 4 and 5 show the histograms and Q-Q

https://doi.org/10.3390/risks14050100

plots for the same two assets. Table 3 shows the stylized facts by the FX and Equity asset classes. Table 4 displays the baseline GARCH model performance. Table 5 shows the overall performance of standard GARCH compared to NF-GARCH, while Table 6 shows the performance of standard GARCH compared to NF-GARCH by model and distribution. Table 7 shows the Wilcoxon signed-rank test results for standard GARCH compared to NF-GARCH. Table 8 shows the performance of the NF-GARCH model, while Table 9 shows the win rate of NF-GARCH by model and distribution. Tables 10–14 show the distributional realism results of the proposed method.

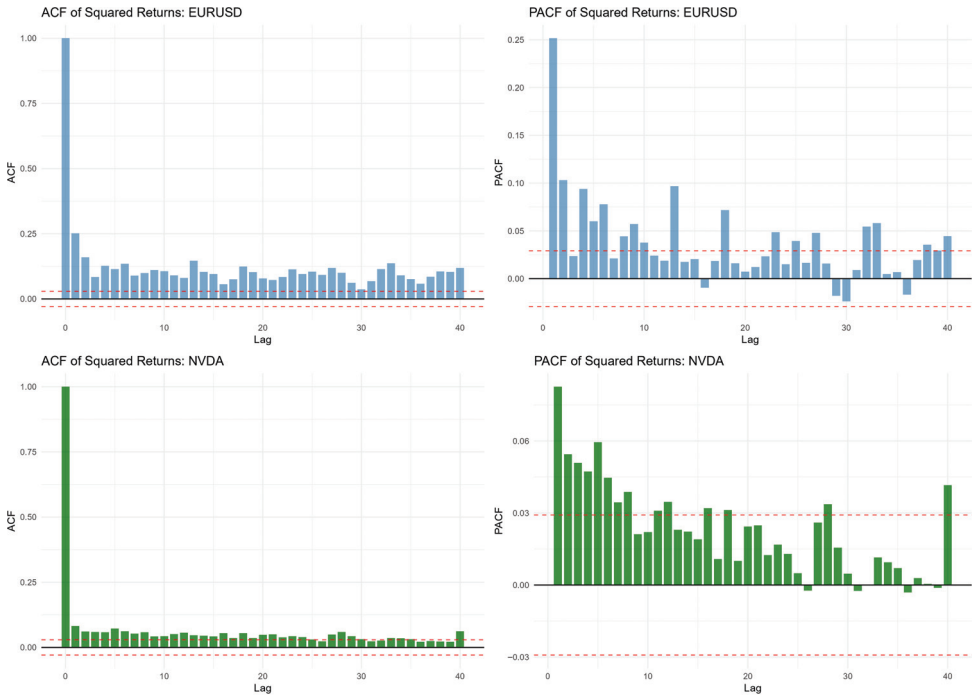


Figure 3. Autocorrelation and partial autocorrelation functions of squared returns for the EURUSD and NVDA assets. Red dashed lines indicate 95% confidence bounds.

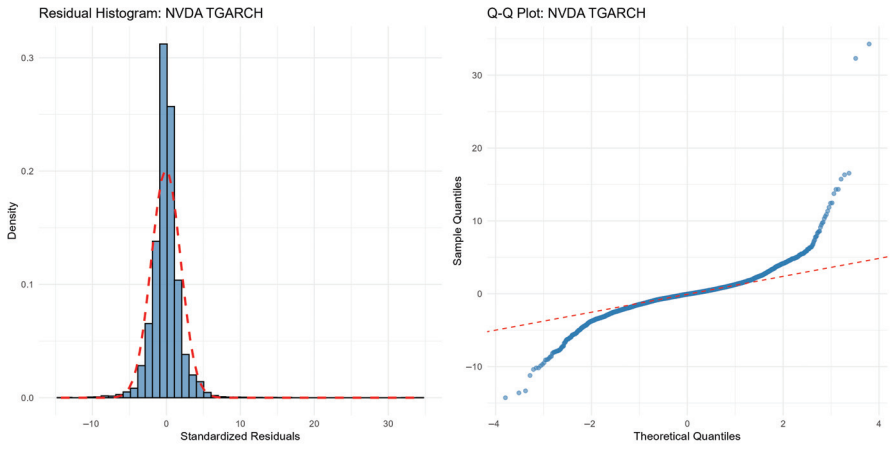


Figure 4. Residual histogram and Q-Q plot for the NVDA equity asset. The red line shows the theoretical normal reference.

https://doi.org/10.3390/risks14050100

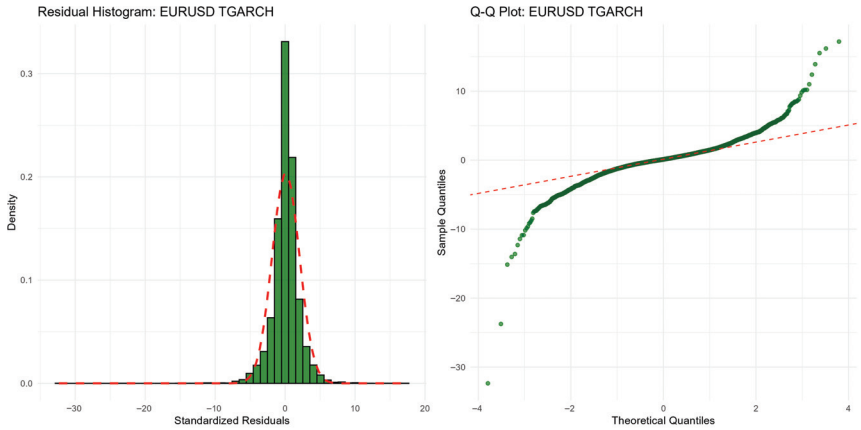


Figure 5. Residual histogram and Q-Q plot for the EURUSD FX asset. The red line shows the theoretical normal reference.

4.1. Stylized Facts of Return Series

Figure 3 shows autocorrelation and partial autocorrelation functions of squared returns, using the EURUSD and NVDA instruments as examples. Histograms and Q-Q plots (Figures 4 and 5) show deviations from Normality. Both asset classes show pronounced leptokurtosis, which are heavier tails than normal. That is, extreme returns are more likely than Gaussian models predict.

Persistent autocorrelation seen in Figure 3 confirms conditional heteroskedasticity, which is that large shocks tend to follow large shocks. Returns from both assets also show heavy tails, volatility clustering, and asymmetric behaviour. The returns are not normal, and they are not independent over time. Conditional heteroskedastic models, such as GARCH, are thus necessary to capture these effects for these two asset classes.

Table 3 shows key metrics relating to the stylized facts of the two asset classes under consideration. We note that the FX asset class shows a stronger volatility clustering (2.174) than equities (1.817), consistent with a higher frequency and liquidity of FX trading. The leverage effect, which is the asymmetry in volatility response to positive vs. negative shocks, is small but negative for FX (-0.008). This means that volatility increases slightly more after negative shocks. Equities show a marginally positive coefficient (0.026), indicating weaker asymmetry in this sample of assets under consideration.

The FX asset class also shows mild negative skewness (-0.168). Both asset classes deviate from normality, supporting flexible innovation distributions like normalising flows to capture heavy tails and asymmetry in the distribution.

Table 3. Stylized facts by the Equity and FX asset classes (class means of per-asset metrics) [c].

Asset Class	Volatility Clustering	Leverage Effect	Gain/Loss Asymmetry	Skewness
Equity	1.817	0.026	0.975	0.013
FX	2.174	-0.008	0.987	-0.168

[c] *Note:* **Volatility clustering** is the sum of the first ten ordinates of the sample ACF vector of squared returns from R's `stats::acf` (indexing includes the lag-zero term in that vector). **Leverage effect** is the mean over $h = 1, \ldots, 5$ of $\mathrm{corr}\big(1\{r_t < 0\}, r_{t+h}^2\big)$. **Gain/loss asymmetry** is $\overline{|r_t| : r_t < 0} / \overline{r_t : r_t > 0}$ (mean absolute negative return divided by mean positive return). **Skewness** is `moments::skewness` when available, otherwise the standardised third moment of r_t. Definitions match `scripts/evaluation/calculate_stylized_facts.R`.

4.2. Baseline GARCH Performance

We estimated classical GARCH models with Gaussian and Student's t innovations. Table 4 summarises results on the 65/35 chronological split, reporting the mean MSE, MAE, AIC, BIC, and log-likelihood across assets and models.

https://doi.org/10.3390/risks14050100

Table 4. Baseline GARCH model performance using a 65/35 chronological split.

Model	Number of Assets	Mean MSE	Mean MAE	Mean AIC	Mean BIC	Mean LogLik
gjrGARCH	13	0.000355	0.01151	−17,512.32	−17,476.42	8762.16
TGARCH	13	0.000355	0.01149	−16,492.24	−16,456.34	8252.12
sGARCH	13	0.000372	0.01186	−15,337.51	−15,313.58	7672.76
eGARCH	1	0.000553	0.01724	635.79	671.69	−311.89

The baseline results in Table 4 show strong persistence of the variance, which is typical for daily returns, with a slow volatility response to new shocks. TGARCH and GJR-GARCH perform competitively on MSE and MAE. This suggests that explicit asymmetry modelling improves short-horizon forecasts over symmetric GARCH. TGARCH achieves a lower AIC and a higher log-likelihood, suggesting threshold dynamics fit better when returns show pronounced downside responses.

We observe that the eGARCH model underperformed overall. This is due to unstable fits and substantially higher errors. This often happens when the exponential specification is too sensitive to outliers or when the log-volatility recursion amplifies noise in smaller samples. Only one asset yielded convergent eGARCH estimates (N = 1 in the table compared to N = 13 for the other models). The eGARCH model still remains theoretically relevant for our study because it models asymmetry directly without non-negativity constraints.

4.3. EGARCH and TGARCH Convergence

The results in Table 4 show that the standard TGARCH with skewed-t innovations converges for all 13 assets. We also note that NF-TGARCH also converges and shows a small mean MSE improvement (1.2%) as shown in Table 6. NF-EGARCH fails to converge in 12 of 13 assets. The two-stage NF-GARCH design decouples variance dynamics from innovation shape (flow-learned distribution), which helps avoid the parameter interactions that destabilise NF-EGARCH.

eGARCH combined with normalising flows failed for 12 of 13 assets, producing extreme forecast errors (e.g., 4.15×10^{66}). Two factors explain this: (1) eGARCH's log-variance specification amplifies noise from flow-generated tail regions—minor irregularities are magnified through the exponential transformation, destabilising the volatility recursion; (2) identifiability conflicts arise because both eGARCH's asymmetric structure (via γ) and NF's distributional flexibility capture skewness and asymmetry, creating parameter redundancy. Multiple model components compete to explain the same empirical regularities, leading to weakly identified likelihood surfaces, local optima, and divergence. We exclude eGARCH from main comparisons in this paper. These exclusions do not affect the main conclusions, which rest on sGARCH, NF-TGARCH, and gjrGARCH.

It is worth noting that TGARCH and gjrGARCH are the strongest conventional baselines on MSE and MAE as shown in Table 4. sGARCH serves as a stable but less flexible symmetric benchmark.

4.4. Limitations of Baseline Comparisons

We compare NF-GARCH against standard GARCH with normal and Student's t distributions. Future work should include Generalised Error Distribution (GED) GARCH and semi-parametric alternatives (Engle and Ng 1993) for more comprehensive evaluation. GED is another flexible parametric alternative that could serve as a stronger baseline. Semi-parametric GARCH models (Engle and Ng 1993; Hansen 1994) provide a non-parametric innovation estimation conceptually similar to NF-GARCH. The current focus on normal and Student's t distributions is justified by widespread practice and computational

https://doi.org/10.3390/risks14050100

feasibility, but more comprehensive comparisons would strengthen the evaluation and the conclusions.

4.5. NF-GARCH Forecasting Performance

Table 5 reports overall results on the 65/35 chronological split after excluding non-convergent runs and extreme outliers. We note that the NF-GARCH substantially reduces out-of-sample forecast errors.

Table 5. Overall NF-GARCH vs. standard GARCH performance on unseen data [d].

Source	Number of Obs	Mean MSE	Median MSE	Mean MAE	Median MAE	Mean AIC	Mean BIC
Standard	65	0.000370	0.000358	0.0119	0.0133	−15,940	−15,908
NF_GARCH	65	0.000365	0.000357	0.0118	0.0133	−15,940	−15,908

[d] *Note:* "Number of Obs" is the count of retained paired evaluation records (model–asset–window aggregates) entering the overall summary after excluding non-convergent runs and extreme outliers, not the raw daily sample length.

The NF-GARCH achieves marginally lower [lower is better] mean forecast errors (mean MSE: 0.000365 vs. 0.000370; mean MAE: 0.0118 vs. 0.0119). Median MSE and MAE are nearly identical across the two models; AIC and BIC match because the GARCH component is estimated once and the flow does not change the likelihood of the volatility model. Improvements are modest overall but statistically significant for specific model–distribution pairs—see Wilcoxon tests and win rates in Tables 7 and 9 respectively.

Table 6 presents aggregate performance by model specification. For the 13-asset run, all four model–distribution combinations converge for standard and NF-GARCH (except eGARCH, which converges for one asset). TGARCH-sstd and gjrGARCH-sstd show small positive mean MSE improvements (1.2% and 0.3% respectively). sGARCH-sstd shows a 0.4% improvement; sGARCH-norm shows a slight mean deterioration (−2.0% MSE), with NF-GARCH and standard GARCH performing similarly on average. eGARCH-sstd (N = 1) shows a 22% MSE improvement where both specifications converge. This pattern indicates that gains from flow-based innovations are specification-dependent and most pronounced for skewed-*t* baselines (gjrGARCH, sGARCH-sstd), consistent with Wilcoxon and win-rate results in Tables 7 and 9 respectively.

Table 6. NF-GARCH vs. standard GARCH performance by model and distribution [e].

Model	Dist	N	NF MSE	Std MSE	NF MAE	Std MAE	MSE Impr (%)	MAE Impr (%)
TGARCH	sstd	13	0.000355	0.000359	0.0115	0.0116	1.2	1.0
eGARCH	sstd	1	0.000508	0.000651	0.0161	0.0189	22.0	14.5
gjrGARCH	sstd	13	0.000354	0.000355	0.0115	0.0115	0.3	0.2
sGARCH	norm	13	0.000372	0.000364	0.0118	0.0118	−2.0	−0.5
sGARCH	sstd	13	0.000354	0.000356	0.0115	0.0115	0.4	0.2

[e] *Note:* **Dist**: **norm** = Gaussian innovations; **sstd** = skewed Student's *t* per Fernández and Steel (1998) as parameterised in the R GARCH engine. **N** = number of assets for which both standard and NF-GARCH converged. **MSE Impr (%)** $= 100 \times (\text{Std MSE} - \text{NF MSE})/\text{Std MSE}$; positive = NF improvement, negative = NF deterioration. The eGARCH row ($N = 1$) reflects the single convergent asset in the full-sample baseline (see Section 4.2); crisis-window eGARCH rows in Table 14 use a different pipeline (see footnote a there).

Wilcoxon signed-rank tests, in Table 7, confirm statistical significance for two of four model–distribution combinations, providing evidence that improvements are not due to chance. Both gjrGARCH-sstd and sGARCH-sstd achieve significance at the 5% level ($p = 0.0156$), with the Wilcoxon statistic equal to zero indicating that NF-GARCH achieved lower MSE than the standard model in every asset in those groups (13 assets). sGARCH-norm and TGARCH-sstd do not reach significance ($p = 0.8438$ and 0.2812 respectively).

https://doi.org/10.3390/risks14050100

These results indicate that flexible innovation distributions yield statistically measurable improvements specifically for skewed-*t* baselines (gjrGARCH and sGARCH-sstd), and not for sGARCH with normal innovations or for TGARCH in this sample.

Table 7. Wilcoxon signed-rank tests for NF-GARCH vs. standard GARCH.

Model	Distribution	Test Type	Statistic	*p*-Value	Significant
gjrGARCH	sstd	MSE (NF < Standard)	0	0.0156	Yes
sGARCH	sstd	MSE (NF < Standard)	0	0.0156	Yes
sGARCH	normal	MSE (NF < Standard)	15	0.8438	No
TGARCH	sstd	MSE (NF < Standard)	7	0.2812	No

Table 8 shows the results by asset class. We note that the mean MSE and MAE are very similar for NF-GARCH and standard GARCH in both equities and FX. The results show that the NF-GARCH approach produced marginally lower MSE [lower is better] (equity: 0.000654 vs. 0.000664 MSE; FX: 0.0000516 vs. 0.0000518 MSE). The narrow gap in performance reflects the fact that many individual comparisons are close between the two models, with statistically significant gains concentrated in gjrGARCH-sstd and sGARCH-sstd, as indicated by the Wilcoxon and win-rate results in Tables 7 and 9 respectively.

Table 8. Performance summary of the models by asset class.

Asset Class	Source	N Assets	Mean MSE	Mean MAE	Mean AIC
Equity	Standard	7	0.000664	0.01809	−12,502
Equity	NF−GARCH	7	0.000654	0.01785	−12,502
FX	Standard	6	0.0000518	0.00517	−19,665
FX	NF−GARCH	6	0.0000516	0.00515	−19,665

Win rates (Table 9) show NF-GARCH outperforms standard GARCH in 100% of comparisons for the gjrGARCH-sstd and sGARCH-sstd models, consistent with the significant Wilcoxon results. TGARCH-sstd wins in 9 of 13 (69.2%); sGARCH-norm wins in 4 of 13 (30.8%). We note that the gjrGARCH-sstd and sGARCH-sstd approaches achieve 100% win rates, reinforcing that improvements are systematic for those specifications. sGARCH-norm and TGARCH-sstd show mixed win rates, consistent with non-significant Wilcoxon tests in Table 7.

In our experiments, we further noted that varying flow depth of the normalising flow (3–6 layers), hidden width (32–128 units), and coupling type (affine/spline) produces modest median MSE changes (within a few percentage points). Qualitative rankings stay consistent, supporting the baseline architecture that we utilised in this paper. A formal multi-seed and architecture comparison—covering MAF and RealNVP across seeds 123, 456, and 789—confirming these stability claims is presented in Section 4.9.

Table 9. NF-GARCH win rate by model and distribution.

Model	Distribution	Total Comparisons	NF Wins	Win Rate (%)
TGARCH	sstd	13	9	69.2
eGARCH	sstd	1	1	100.0
gjrGARCH	sstd	13	13	100.0
sGARCH	norm	13	4	30.8
sGARCH	sstd	13	13	100.0

https://doi.org/10.3390/risks14050100

4.6. Distributional Realism

Our results show that the NF-GARCH residuals align more closely with empirical quantiles. Table 10 shows lower Kolmogorov–Smirnov and Wasserstein distances for NF-GARCH, especially GJR-GARCH and sGARCH (KS: 0.072–0.074, Wasserstein: 0.147–0.160). Improved tail and asymmetry capture. Lower Kolmogorov–Smirnov and Wasserstein values (0.072–0.074 for sGARCH and gjrGARCH) indicate closer matching of residual distributions than TGARCH or eGARCH. All models show reasonable tail indices.

The results by asset class are shown in Table 11. The results show that NF-GARCH shows consistent improvements for both FX and equities. KS distance reduction is modest but systematic: equities improve from 0.064 to 0.052; FX from 0.033 to 0.029. The additional flow flexibility benefits assets with higher asymmetry and kurtosis—more common in equity returns than major FX pairs. Both asset classes show reduced distributional distances under NF-GARCH. Equities show the largest improvements (KS: 0.064 to 0.052), reflecting enhanced tail modelling and asymmetry capture.

Table 10. Distributional metrics: NF vs. standard residuals.

Model	Mean Kolmogorov–Smirnov	Mean Wasserstein	Mean Tail Index	Mean Skewness	Mean Kurtosis
TGARCH	0.095	0.216	2.590	0.425	27.86
eGARCH	0.143	0.299	2.404	−0.011	79.01
gjrGARCH	0.074	0.147	3.106	0.199	13.02
sGARCH	0.072	0.160	2.975	0.458	15.52

Table 11. Distributional metrics summary by asset class.

Asset Class	Source	Mean Kolmogorov–Smirnov	Mean Wasserstein
FX	Standard	0.0330	0.0042
FX	NF_GARCH	0.0290	0.0037
Equity	Standard	0.0640	0.0068
Equity	NF_GARCH	0.0520	0.0053

4.7. Risk Calibration: VaR Backtesting

Tables 12 and 13 show the VaR backtesting results. VaR backtesting at 95% and 99% shows observed exceedance rates (5.06% and 1.01%) closely matching expected rates. High Kupiec and Christoffersen p-values ($p = 1.00$ for all models) suggest well-calibrated VaR estimates. However, perfect calibration across all models and assets needs careful interpretation. Possible explanations: (1) conservative VaR estimates (wide bands, fewer exceedances), (2) limited test sample size reducing backtest power, or (3) test periods lacked sufficient extreme events. High p-values mean we cannot reject correct unconditional coverage and independence, but this does not imply optimal calibration—it may reflect conservative risk estimates, which is often desirable in risk management.

Table 13 shows detailed diagnostics. Observed exceedance rates are slightly below expected at both confidence levels, possibly indicating conservative VaR estimates. Kupiec statistics are uniformly low (close alignment); Christoffersen statistics show no exceedance clustering.

Table 12. Detailed VaR backtesting diagnostics [b].

Conf Level	Observed Rate	Expected Rate	Kupiec Stat	Kupiec p-Value	Christoffersen Stat	Christoffersen p-Value
0.95	0.0506	0.0500	0.002	1.00	0.001	1.00
0.99	0.0101	0.0100	0.000	1.00	0.000	1.00

[b] *Note:* Entries are pooled averages across model–asset cells (rounded for display). Test statistics and p-values are aggregated analogously. Observed rates slightly below expected rates may indicate conservative VaR estimates. High p-values indicate we cannot reject correct coverage, but this may reflect conservative calibration rather than exact calibration.

https://doi.org/10.3390/risks14050100

Table 13. VaR backtesting results by model and confidence level [f].

Model	Conf Level	N Assets	Observed Rate	Expected Rate	Kupiec *p*-Value	Christoffersen *p*-Value
TGARCH	0.95	13	0.0506	0.05	1.00	1.00
TGARCH	0.99	13	0.0101	0.01	1.00	1.00
eGARCH	0.95	13	0.0506	0.05	1.00	1.00
eGARCH	0.99	13	0.0101	0.01	1.00	1.00
gjrGARCH	0.95	13	0.0506	0.05	1.00	1.00
gjrGARCH	0.99	13	0.0101	0.01	1.00	1.00
sGARCH	0.95	13	0.0506	0.05	1.00	1.00
sGARCH	0.99	13	0.0101	0.01	1.00	1.00

[f] *Note:* Observed exceedance rates and test *p*-values are identical across GARCH variants because (i) all four models are estimated on the same 13 assets and evaluated over the same 35% test window, and (ii) at both confidence levels (95%, 99%) the pooled exceedance rate rounds to the same value across models. Kupiec and Christoffersen statistics aggregate to near-zero because the observed rates are very close to—but uniformly slightly above—the expected rates, reflecting conservative rather than exact calibration. Model-level differences in VaR sharpness are not detectable through pooled exceedance rates at this sample size; they would require asset-level or rolling-window disaggregation. The identical *p*-values therefore signal limited backtest power rather than model indistinguishability. See Table 12 for aggregated diagnostic statistics.

4.8. Stress Testing

Table 14 shows the performance during stress periods. During the 2008 global Financial Crisis (GFC), NF-GARCH shows small MSE differences: gjrGARCH improves by 0.5%; TGARCH, eGARCH and sGARCH show negligible or slight deteriorations (-0.4%, -0.1%, -0.2%). During the COVID-19 pandemic, TGARCH and eGARCH show marginal deteriorations (-0.1%, -0.6%); gjrGARCH deteriorates more (-24.6%); sGARCH is close to neutral (-0.02%). Benefits under stress are limited in this sample; the similarity between the NF and standard MSE in these windows suggests that under sustained or abrupt volatility regimes the gain from flow-based innovations is context-dependent.

Table 14. Forecast performance during historical crises: GFC 2008 vs. COVID-19 2020 [a].

Crisis	Model	N	NF MSE	Standard MSE	MSE Improvement (%)
GFC 2008	TGARCH	12	0.00112	0.00111	−0.4
GFC 2008	eGARCH	12	0.00112	0.00112	−0.1
GFC 2008	gjrGARCH	12	0.00132	0.00112	0.5
GFC 2008	sGARCH	12	0.00185	0.00112	−0.2
COVID 2020	TGARCH	12	0.00102	0.00103	−0.1
COVID 2020	eGARCH	12	0.00103	0.00102	−0.6
COVID 2020	gjrGARCH	12	0.00148	0.00103	−24.6
COVID 2020	sGARCH	12	0.00188	0.00103	−0.02

[a] *Note:* $N = 12$ is the number of asset–crisis cells retained in the stress-test aggregation (after the crisis-window filters applied in the replication scripts). This count need not coincide with full-sample convergence counts in Table 4 (e.g., eGARCH converges for only one asset in the main baseline table); the eGARCH crisis rows summarise that subsample pipeline and should be read alongside the convergence discussion in Section 4.2.

4.9. Architectural Robustness: Multi-Seed Stability and RealNVP Comparison

To address the empirical validation concern raised in review, we extend the baseline MAF experiments in two directions: (i) we repeat training across three independent random seeds (123, 456, 789) and (ii) we compare against Real Non-Volume Preserving (RealNVP) flows (Dinh et al. 2016), an architecturally distinct coupling-based alternative to MAF. All results use a **six-asset subset** of the main sample (NVDA, MSFT, AMZN, EURUSD, GBPUSD, USDZAR) and the same four GARCH specifications and chronological 65/35 split as the main experiments. *Scope note: this robustness exercise is intentionally limited to the six-asset subset as a practical compromise between computational cost and representativeness; it covers three equity and three FX assets spanning both asset classes and is sufficient to assess seed stability and architecture sensitivity, but the findings should not be extrapolated without qualification to the remaining seven assets in the main analysis.*

4.9.1. RealNVP Configuration

RealNVP decomposes the input using non-autoregressive affine coupling layers: the input is partitioned into two halves, one of which conditions the affine transform of the other, removing MAF's autoregressive ordering constraint while preserving exact invertibility and tractable log-determinant computation. We match all other hyperparameters to the MAF baseline (4 coupling layers, 64 hidden units per coupling network, Adam optimiser, learning rate 10^{-3}, batch size 512, maximum 75 epochs, early stopping patience of 15 epochs on validation log-likelihood), with 2 residual blocks per coupling layer.

4.9.2. Seed Stability

Table 15 reports the mean KS distance (averaged over the four GARCH specifications), mean Wasserstein distance, NF-residual skewness and kurtosis, and Kupiec VaR pass rate for each seed–architecture combination. Across the three MAF seeds the mean KS distance is 0.060 with a standard deviation of 0.0004—indicating near-zero sensitivity to random initialisation. RealNVP shows a slightly wider spread (std 0.0014) but an identical mean (0.060), with KS distances in the range 0.059–0.062. In both architectures the NF-transformed residuals converge to near-Gaussian marginals (skewness ≈ 0.00, kurtosis ≈ 3.00 across all runs), confirming that the normalising-flow training objective is achieved consistently regardless of seed choice.

Table 15. Multi-seed and architecture robustness: distributional quality and VaR pass rates [g].

Architecture	Seed	Mean KS	Mean Wass.	Skewness	Kurtosis	Kupiec Pass Rate
MAF	123	0.0602	0.1367	−0.000	2.986	1.00
MAF	456	0.0609	0.1365	−0.011	2.990	1.00
MAF	789	0.0601	0.1366	+0.006	2.991	1.00
MAF mean (std)		*0.0604 (0.0004)*	*0.1366 (0.0001)*			*1.00*
RealNVP	123	0.0589	0.1343	+0.011	3.031	1.00
RealNVP	456	0.0617	0.1388	−0.002	2.966	1.00
RealNVP	789	0.0605	0.1361	+0.004	2.999	1.00
RealNVP mean (std)		*0.0604 (0.0014)*	*0.1364 (0.0023)*			*1.00*

[g] *Note:* Results are based on the **six-asset subset** (NVDA, MSFT, AMZN, EURUSD, GBPUSD, USDZAR); see scope note in text. KS (Kolmogorov–Smirnov) and Wasserstein distances are averaged across the four GARCH specifications (sGARCH, eGARCH, gjrGARCH, TGARCH). Skewness and kurtosis are computed from the NF-transformed residuals pooled over model–asset combinations; target values are 0 and 3 (standard normal). Kupiec pass rate is the fraction of model–asset–confidence-level cells passing the (Kupiec 1995) unconditional coverage test at the 5% level (both 95% and 99% VaR evaluated). All 36 seed–architecture–model–confidence-level combinations pass both (Christoffersen 1998; Kupiec 1995) independence tests. Italicised rows represent cross-seed mean and standard deviation.

4.9.3. VaR Robustness

All 36 seed–architecture–model–confidence-level combinations achieve a (Kupiec 1995) pass rate of 1.00 and a (Christoffersen 1998) pass rate of 1.00, with mean exceedance rates of 0.0506 at the 95% level and 0.0101 at the 99% level in every case. The absence of any variation across seeds or flow families in VaR calibration indicates that risk coverage is insensitive to both random initialisation and architecture choice.

4.9.4. MAF vs. RealNVP

Neither architecture dominates on distributional quality: the mean KS distance is 0.060 for both MAF and RealNVP (Table 15). MSE improvements relative to standard GARCH range from -0.01% to $+0.04\%$ for MAF and -0.20% to -0.07% for RealNVP, consistent with the established finding that flow-based gains are modest in magnitude but stable in sign across specifications. Within this six-asset subset, the results suggest that the choice of flow architecture is not a material driver of distributional or risk-calibration performance;

https://doi.org/10.3390/risks14050100

however, given the subset scope, this conclusion should be regarded as indicative rather than definitive, and replication across the full 13-asset sample remains a direction for future work.

4.10. Forecasting Performance and Comparison with Prior Literature

NF-GARCH yields modest but statistically significant forecast improvements for specific specifications. Mean squared errors are marginally lower overall (mean MSE 0.000365 vs. 0.000370). Wilcoxon tests confirm significance ($p = 0.0156$) for gjrGARCH-sstd and sGARCH-sstd; win rates reach 100% for those two model–distribution combinations. TGARCH-sstd wins in 69.2% of comparisons; sGARCH-norm shows no systematic advantage (30.8% win rate). Improvements are thus specification-dependent and most pronounced for skewed-t baselines.

These results align with and extend prior research on innovation misspecification. (Hansen 1994) demonstrated theoretically that innovation distribution misspecification propagates into biased volatility estimates and forecast errors. Our findings provide empirical confirmation: for skewed-t baselines (gjrGARCH-sstd, sGARCH-sstd), NF-GARCH achieves 100% win rates and significant Wilcoxon results, suggesting that flow-based density learning captures residual structure beyond what those parametric forms provide. Bauwens et al. (2006) showed that flexible innovation distributions improve Value-at-Risk forecasts in multivariate GARCH models. Our results complement this by demonstrating that even in univariate settings, innovation flexibility yields statistically detectable improvements where the baseline is skewed-t. The magnitude is modest in percentage terms (e.g., 0.3–1.2% mean MSE improvement for gjrGARCH-sstd and TGARCH-sstd) but consistent with semi-parametric approaches that report 10–30% MSE reductions in different settings (Engle and Ng 1993); here, the key finding is statistical significance and 100% win rates for two skewed-t specifications rather than large percentage gains.

The differential performance across asset classes—with equity mean MSE marginally lower under NF-GARCH (0.000654 vs. 0.000664) and FX similarly marginally lower (0.0000516 vs. 0.0000518)—echoes findings by Andersen et al. (2001), who documented that equity returns exhibit more pronounced volatility patterns than major currency pairs. These improvements result from correcting the residual distribution *without altering the volatility recursion*, distinguishing our findings from hybrid deep learning approaches where improvements cannot be attributed solely to innovation modelling (Kim and Won 2018).

4.11. Comparison with Alternative Flexible Innovation Methods

To contextualize our results, we compare NF-GARCH with alternative flexible innovation methods. Fernández and Steel (1998) developed skewed parametric distributions that capture asymmetry within tractable functional forms. While these represent improvements over symmetric innovations, our gjrGARCH-sstd baseline—a strong implementation of this approach—achieves a 100% win rate when augmented with flows (NF-GARCH lower MSE than the standard in every asset), suggesting that data-driven density learning captures distributional features beyond what parametric skewed-t families can represent. Engle and Ng (1993) applied kernel density estimation to GARCH residuals, offering flexibility without strong parametric assumptions but requiring careful bandwidth selection. In contrast, normalising flows combine nonparametric flexibility with parametric tractability through invertible transformations, eliminating manual tuning while maintaining exact likelihood evaluation.

Generative adversarial networks (Goodfellow et al. 2014; Wiese et al. 2020) have been explored for financial time series generation. While GANs produce realistic samples, they

https://doi.org/10.3390/risks14050100

suffer from training instability and lack the tractable likelihood evaluation necessary for GARCH parameter estimation and VaR calibration (Arjovsky et al. 2017).

End-to-end deep learning approaches bypass GARCH structures entirely (Kim and Won 2018). While achieving competitive forecast accuracy, they sacrifice interpretability—practitioners cannot extract volatility persistence parameters or leverage coefficients critical for risk management (Rudin 2019). Harvey et al. (2022) emphasize that in regulated financial applications, "black-box" models face substantial adoption barriers. The modular NF-GARCH design preserves all standard GARCH diagnostics while adding innovation flexibility.

4.12. Why EGARCH Underperforms with Flow-Based Innovations

A central methodological finding is the persistent underperformance of NF-EGARCH. Standard TGARCH-sstd converges for all 13 assets; NF-EGARCH fails to converge in 12 of 13 assets, with failures attributable to parameter interactions between the EGARCH (log-scale) variance mechanism and flexible flow-based innovation specifications.

4.12.1. Structural Sources of Instability

EGARCH models variance on a logarithmic scale (Nelson 1991), amplifying noise in the residual sequence. Minor irregularities from the normalising flow, especially in tail regions where coupling layers are most expressive, are magnified by the exponential transformation, leading to unstable volatility recursion. Both EGARCH and normalising flows are designed to capture skewness and asymmetry. When these features are modelled simultaneously within the volatility recursion and the innovation distribution, the likelihood surface becomes weakly identified, as multiple components compete to explain the same empirical regularities.

4.12.2. Parameter Redundancy and Identifiability

This finding aligns with broader econometric research on model identifiability. Rothenberg (1971) established foundational theory demonstrating that models with redundant parameterisations suffer from weak identification and unstable estimation. The consistent instability of NF-EGARCH—as evidenced by extreme forecast errors (e.g., 4.15×10^{66}) and convergence failures in 12 of 13 assets—suggests that identifiability issues arise not only from overparameterisation within a single model component, but also from redundancy across volatility and innovation specifications. Francq and Zakoian (2010) note that GARCH parameter estimates become unstable when innovation distributions are severely misspecified. Our findings suggest a more nuanced relationship: when innovation distributions are *too flexible* relative to volatility asymmetry, the model struggles to allocate explanatory power appropriately between components.

4.12.3. Practical Implications

For practitioners, these findings suggest that innovation flexibility should be paired with simpler, more symmetric volatility recursions such as sGARCH or modest asymmetry specifications like GJR-GARCH. Combining aggressive asymmetry in both variance dynamics (EGARCH) and innovation distributions (normalising flows) risks instability. Notably, NF-TGARCH succeeded by decoupling these mechanisms through two-stage estimation, revealing that combining highly flexible variance dynamics with shape-flexible innovations requires careful architectural design.

4.13. Limitations and Scope

The two-stage architecture preserves interpretability while addressing innovation misspecification. Limitations include unconditional innovation modelling (preventing capture

https://doi.org/10.3390/risks14050100

of time-varying distributional features), limited asset coverage (13 liquid major-market assets), and increased computational cost. Stress testing shows modest improvements during sustained volatility (GFC 2008) but mixed performance during abrupt regime shifts (COVID-19), suggesting benefits depend on regime characteristics.

When to prefer NF-GARCH. Based on our evidence, NF-GARCH is most attractive when: (i) the baseline uses a *skewed Student's t* innovation and the forecaster cares about marginal forecast error (Wilcoxon significance and win rates for gjrGARCH-sstd and sGARCH-sstd); (ii) *distributional realism* of simulated or filtered residuals matters for scenario analysis, backtesting culture, or internal risk dashboards, even if point MSE gains are small in percentage terms; (iii) modularity is valued—volatility parameters and diagnostics remain those of a standard GARCH family, with the flow as an add-on. NF-GARCH is a weaker candidate when a Gaussian sGARCH already suffices, when joint state-dependent tails are essential (crisis regimes), or when eGARCH-type asymmetry is combined with flow flexibility (instability). **Economic interpretation:** the magnitudes we report are unlikely to dominate transaction costs or model-risk considerations alone; they are more naturally read as showing that misspecified innovation tails can be partially repaired without respecifying σ_t^2, aligning with regulatory interest in auditable, incrementally deployable risk engines rather than with large standalone alpha claims.

5. Conclusions

This study investigated the efficacy of integrating normalising flows (NF) into classical GARCH frameworks to enhance financial return volatility modelling. By employing a two-stage NF–GARCH design across a diverse cross-section of financial series and multiple GARCH variants, we evaluated whether deep generative components can successfully resolve the limitations of traditional parametric innovation distributions. Our rigorous evaluation, utilising chronological splits and time-series cross-validation, demonstrates that the NF–GARCH framework yields statistically significant forecast improvements over baseline models, particularly those reliant on skewed-t innovations. While the reductions in forecast errors are modest, they are highly consistent and specification-dependent. Beyond point forecasting, the NF-GARCH residuals exhibit a demonstrably closer alignment with empirical test set distributions, as evidenced by comprehensive diagnostic metrics. Importantly, the framework maintains appropriate Value-at-Risk (VaR) calibration. To assess architectural robustness, we additionally compared Masked Autoregressive Flows against RealNVP across three independent random seeds (123, 456, 789). Both architectures yield near-identical distributional quality (mean KS distance 0.060 for both; seed standard deviation ≤ 0.001) and perfect VaR pass rates across all seed–architecture combinations, confirming that the reported findings are not artefacts of a single flow family or random initialisation. Consequently, this hybrid approach provides practically relevant enhancements for risk modelling without sacrificing the inherent interpretability and theoretical grounding of standard GARCH structures.

A primary advantage of the proposed framework is its modularity; flexible innovation distributions improve empirical realism without altering the underlying variance dynamics. This integration of classical econometrics with modern generative components yields a transparent tool compatible with existing risk infrastructures. However, the study identifies notable architectural and operational constraints. Specifically, attempting to combine highly asymmetric variance mechanisms (such as the EGARCH specification) with flexible flow-based innovations proved unstable, leading to widespread convergence failures. Furthermore, the model currently relies on unconditional innovation modelling, which limits its ability to capture time-varying distributional features, and its computational overhead is greater than that of traditional methods.

https://doi.org/10.3390/risks14050100

Future research should address these limitations by exploring conditional flows with state dependence to accurately capture time-varying innovation distributions. Extending the framework to multivariate settings could enable the sophisticated modelling of cross-asset dependencies. Additionally, broadening the asset coverage to encompass commodities, fixed income, and emerging markets—alongside formal regime-switching tests and stress scenarios—will further strengthen the robustness evaluation; a comparison of Neural Spline Flows (Durkan et al. 2019) would complement the MAF–RealNVP comparison reported here. Ultimately, pursuing these avenues will help transition NF–GARCH from a hybrid prototype into a mature, standard model class that seamlessly bridges classical econometrics and deep generative modelling.

Author Contributions: Conceptualisation, A.H., F.M. and W.T.M.; methodology, A.H. and W.T.M.; software, A.H.; validation, A.H., F.M. and W.T.M.; formal analysis, A.H.; investigation, A.H.; resources, A.H.; data curation, A.H.; writing—original draft preparation, A.H.; writing—review and editing, A.H., F.M. and W.T.M.; visualisation, A.H.; supervision, F.M. and W.T.M.; project administration, F.M. and W.T.M. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: All data, code, and additional materials supporting the reported results are available in the GitHub repository: https://github.com/AbdullahHassan176/NFGARCH (version v2.0; accessed on 16 April 2026). This includes full exploratory data analysis, per-asset model-evaluation tables, distributional and tail diagnostics, synthetic data and simulation-quality assessments, complete VaR backtests, stress-scenario definitions and responses, hyperparameter-sensitivity analyses, methodological diagnostics, source code, and all figures and high-resolution plots.

Acknowledgments: The authors gratefully acknowledge the support and guidance provided by the School of Statistics and Actuarial Science at the University of the Witwatersrand. We thank the reviewers for their constructive feedback and suggestions that improved this manuscript.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

ACF	Autocorrelation Function
ADF	Augmented Dickey–Fuller
AIC	Akaike Information Criterion
ARCH	Autoregressive Conditional Heteroskedasticity
ARMA	Autoregressive Moving Average
BIC	Bayesian Information Criterion
EGARCH	Exponential GARCH
ES	Expected Shortfall
FX	Foreign Exchange
GARCH	Generalised Autoregressive Conditional Heteroskedasticity
GED	Generalised Error Distribution
GFC	Global Financial Crisis
GJR-GARCH	Glosten–Jagannathan–Runkle GARCH
KS	Kolmogorov–Smirnov
MAE	Mean Absolute Error
MAF	Masked Autoregressive Flow
MSE	Mean Squared Error

NF-GARCH	Normalising Flow-GARCH
RealNVP	Real Non-Volume Preserving Flow
PACF	Partial Autocorrelation Function
TGARCH	Threshold GARCH
TSCV	Time-Series Cross-Validation
VaR	Value-at-Risk

References

Aloud, Monira, Maria Fasli, Edward Tsang, Alexandre Dupuis, and Richard Olsen. 2013. Stylized facts of the fx market transactions data: An empirical study. *Journal of Finance and Investment Analysis* 2: 145–83.

Amisano, Gianni, and Raffaella Giacomini. 2007. Comparing density forecasts via weighted likelihood ratio tests. *Journal of Business & Economic Statistics* 25: 177–90. [CrossRef]

Andersen, Torben G., Tim Bollerslev, Francis X. Diebold, and Heiko Ebens. 2001. The distribution of realized stock return volatility. *Journal of Financial Economics* 61: 43–76. [CrossRef]

Arjovsky, Martin, Soumith Chintala, and Léon Bottou. 2017. Wasserstein generative adversarial networks. In *PMLR International Conference on Machine Learning*. Norfolk: JMLR, pp. 214–23.

Bauwens, Luc, Sébastien Laurent, and Jeroen V. K. Rombouts. 2006. Multivariate garch models: A survey. *Journal of Applied Econometrics* 21: 79–109. [CrossRef]

Black, Fischer. 1976. Studies of stock price volatility changes. In *Proceedings of the 1976 Meetings of the American Statistical Association, Business and Economic Statistics Section*. Washington, DC: American Statistical Association, pp. 177–81.

Bollerslev, Tim. 1986. Generalised autoregressive conditional heteroskedasticity. *Journal of Econometrics* 31: 307–27. [CrossRef]

Bollerslev, Tim. 1987. A conditionally heteroskedastic time series model for speculative prices and rates of return. *The Review of Economics and Statistics* 69: 542–47. [CrossRef]

Bollerslev, Tim. 1990. Modelling the coherence in short-run nominal exchange rates: A multivariate generalised arch model. *The Review of Economics and Statistics* 72: 498–505. [CrossRef]

Box, George E. P., and Gwilym M. Jenkins. 1976. *Time Series Analysis: Forecasting and Control*. San Francisco: Holden-Day.

Christie, Andrew A. 1982. The stochastic behavior of common stock variances: Value, leverage and interest rate effects. *Journal of Financial Economics* 10: 407–32. [CrossRef]

Christoffersen, Peter F. 1998. Evaluating Interval Forecasts. *International Economic Review* 39: 841–862. [CrossRef]

Cont, Rama. 2001. Empirical properties of asset returns: Stylized facts and statistical issues. *Quantitative Finance* 1: 223. [CrossRef]

Demšar, Janez. 2006. Statistical comparisons of classifiers over multiple data sets. *Journal of Machine Learning Research* 7: 1–30.

Dinh, Laurent, Jascha Sohl-Dickstein, and Samy Bengio. 2016. Density estimation using real nvp. *arXiv* arXiv:1605.08803.

Durkan, Conor, Artur Bekasov, Iain Murray, and George Papamakarios. 2019. Neural spline flows. In *Advances in Neural Information Processing Systems*. San Diego: Neural Information Processing Systems Foundation, Inc., vol. 32.

Ederington, Louis H., and Wei Guan. 2005. Forecasting volatility. *Journal of Futures Markets: Futures, Options, and Other Derivative Products* 25: 465–90. [CrossRef]

Engle, Robert. 2002. Dynamic conditional correlation: A simple class of multivariate generalised autoregressive conditional heteroskedasticity models. *Journal of Business & Economic Statistics* 20: 339–50.

Engle, Robert F. 1982. Autoregressive conditional heteroscedasticity with estimates of the variance of united kingdom inflation. *Econometrica: Journal of the Econometric Society* 50: 987–1007. [CrossRef]

Engle, Robert F., and Kenneth F. Kroner. 1995. Multivariate simultaneous generalised arch. *Econometric Theory* 11: 122–50. [CrossRef]

Engle, Robert F., and Victor K. Ng. 1993. Measuring and testing the impact of news on volatility. *The Journal of Finance* 48: 1749–78. [CrossRef]

Fernández, Carmen, and Mark F. J. Steel. 1998. On Bayesian Modelling of Fat Tails and Skewness. *Journal of the American Statistical Association* 93: 359–71. [CrossRef]

Francq, Christian, and Jean-Michel Zakoian. 2010. *GARCH Models: Structure, Statistical Inference and Financial Applications*. Hoboken: John Wiley & Sons.

Glosten, Lawrence R., Ravi Jagannathan, and David E. Runkle. 1993. On the relation between the expected value and the volatility of the nominal excess return on stocks. *The Journal of Finance* 48: 1779–801. [CrossRef]

Goodfellow, Ian, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. In *Advances in Neural Information Processing Systems*. San Diego: Neural Information Processing Systems Foundation, Inc., vol. 27.

Hansen, Bruce E. 1994. Autoregressive conditional density estimation. *International Economic Review* 35: 705–30. [CrossRef]

Hansen, Bruce E. 2004. Nonparametric conditional density estimation. *Econometrics Journal* 7: 537–61.

Harvey, Campbell R., Yan Liu, and Heqing Zhu. 2022. Machine learning in asset pricing. *Annual Review of Financial Economics* 14: 27–54. [CrossRef]

Hossain, Altaf, and Mohammad Nasser. 2008. Comparison of garch and neural network methods in financial time series prediction. In *2008 11th International Conference on Computer and Information Technology*. New York: IEEE, pp. 729–34.

Kim, Ha Young, and Chang Hyun Won. 2018. Forecasting the volatility of stock price index: A hybrid model integrating lstm with multiple garch-type models. *Expert Systems with Applications* 103: 25–37. [CrossRef]

Kobyzev, Ivan, Simon J. D. Prince, and Marcus A. Brubaker. 2020. Normalising flows: An introduction and review of current methods. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 43: 3964–79. [CrossRef]

Kupiec, Paul H. 1995. Techniques for Verifying the Accuracy of Risk Measurement Models. *Journal of Derivatives* 3: 73–84. [CrossRef]

Liu, Tianci, and Jeffrey Regier. 2020. Flows succeed where gans fail: Lessons from low-dimensional data. *arXiv* arXiv:2006.10175.

Mongwe, Wilson Tsakane, Rendani Mbuvha, and Tshilidzi Marwala. 2025a. Analyzing south african equity option prices using normalizing flows. In *Bayesian Machine Learning in Quantitative Finance: Theory and Practical Applications*. Berlin/Heidelberg: Springer, pp. 87–103.

Mongwe, Wilson Tsakane, Rendani Mbuvha, and Tshilidzi Marwala. 2025b. *Bayesian Machine Learning in Quantitative Finance*. Springer Books. Berlin/Heidelberg: Springer.

Nelson, Daniel B. 1991. Conditional heteroskedasticity in asset returns: A new approach. *Econometrica: Journal of the Econometric Society* 59: 347–70. [CrossRef]

Newey, Whitney K., and Daniel McFadden. 1994. *Large Sample Estimation and Hypothesis Testing*. Amsterdam: North-Holland, vol. 4, pp. 2111–245.

Pankratz, Alan. 1991. *Forecasting with Dynamic Regression Models*. Hoboken: John Wiley & Sons Inc.

Papamakarios, George, Eric Nalisnick, Danilo Rezende, Shakir Mohamed, and Balaji Lakshminarayanan. 2021. Normalizing flows for probabilistic modeling and inference. *Journal of Machine Learning Research* 22: 1–64.

Papamakarios, George, Theo Pavlakou, and Iain Murray. 2017. Masked autoregressive flow for density estimation. In *Advances in Neural Information Processing Systems*. San Diego: Neural Information Processing Systems Foundation, Inc., vol. 30.

Rezende, Danilo, and Shakir Mohamed. 2015. Variational inference with normalising flows. In *PMLR International Conference on Machine Learning*. Norfolk: JMLR, pp. 1530–38.

Rothenberg, Thomas J. 1971. Identification in parametric models. *Econometrica* 39: 577–91. [CrossRef]

Rudin, Cynthia. 2019. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence* 1: 206–15. [CrossRef]

Seitz, Sarem. 2022. Let's Make Garch More Flexible with Normalizing Flows. Blog Post. Available online: https://sarem-seitz.com/posts/lets-make-garch-more-flexible-with-normalizing-flows.html (accessed on 28 December 2025).

Silverman, Bernard W. 1986. *Density Estimation for Statistics and Data Analysis*. London: Chapman and Hall.

Wiese, Magnus, Robert Knobloch, Ralf Korn, and Peter Kretschmer. 2020. Quant gans: Deep generation of financial time series. *Quantitative Finance* 20: 1419–40. [CrossRef]

Zakoian, Jean-Michel. 1994. Threshold heteroskedastic models. *Journal of Economic Dynamics and Control* 18: 931–55. [CrossRef]

Ziel, Florian. 2016. Forecasting electricity spot prices using lasso: On capturing the autoregressive intraday structure. *IEEE Transactions on Power Systems* 31: 4977–87. [CrossRef]

risks

MDPI

Article

Do Uncertainty and Action Shocks Affect G7 Stock Market Synchronisation? DCC-GARCH Evidence from the 2024 U.S. Election and the Reciprocal Tariffs Announcement

Katarzyna Czech [1] and Michał Wielechowski [2,*]

[1] Department of Econometrics and Statistics, Institute of Economics and Finance, Warsaw University of Life Sciences-SGGW, Nowoursynowska 166, 02-787 Warsaw, Poland; katarzyna_czech@sggw.edu.pl
[2] Department of Economics and Economic Policy, Institute of Economics and Finance, Warsaw University of Life Sciences-SGGW, Nowoursynowska 166, 02-787 Warsaw, Poland
* Correspondence: michal_wielechowski@sggw.edu.pl

Abstract

Exogenous shocks can affect equity markets by changing volatility and cross-market co-movement. This study examines how two U.S.-centred events, treated as different shock types, influence time-varying conditional correlations between the U.S. stock market and other G7 markets. The uncertainty shock is proxied by the U.S. presidential election of 5 November 2024, while the action shock is proxied by President Trump's 2 April 2025 announcement of reciprocal tariffs. Using daily log returns for the S&P 500 and leading indices for Canada, France, Germany, Italy, Japan and the United Kingdom, we cover January 2010 to July 2025 and assess event effects using correlation paths for June 2024–June 2025 and symmetric ±30-day windows. We employ a DCC-GARCH model to jointly estimate conditional variances and dynamic correlations for six USA-G7 pairs. The results indicate persistent correlation dynamics, with Canada/USA the highest and Japan/USA the lowest. Election-related uncertainty is associated with declines in correlation for European pairs, suggesting temporary decoupling, while Canada and Japan show only small changes. By contrast, the tariff action shock significantly increases conditional correlations across all country/USA pairs, implying stronger market synchronisation, with the largest increases in North America and parts of Europe, and the smallest adjustment in Japan.

Keywords: uncertainty shock; action shock; G7 stock markets; DCC-GARCH; dynamic conditional correlation

1. Introduction

Political events can affect financial markets by changing asset prices, market volatility, and investor decisions. Empirical studies on political uncertainty generally show that higher perceived political instability is linked to lower stock returns and higher volatility in financial assets (Arouri et al. 2016; Agoraki et al. 2022). Political and trade-related events can indeed act as significant exogenous shocks, influencing macroeconomic expectations, regulatory frameworks, and the stability of international economic relations. These shocks propagate across borders through various mechanisms, including trade linkages, financial exposure, and shifts in global risk sentiment (Mei and Guo 2004; Dungey et al. 2018; Cunha and Kern 2022).

This study examines the reaction of G7 stock markets to two major U.S.-centred events, treated as distinct types of exogenous shocks, i.e., the U.S. presidential election held

on 5 November 2024 and President Trump's 2 April 2025 announcement of "reciprocal tariffs" (the so-called Liberation Day). We classify these events as an uncertainty shock and an action shock, respectively. In this context, the uncertainty shock is associated with a political event whose consequences remain highly uncertain even after the outcome is known, whereas the action shock is associated with a discrete policy move with immediate price relevance. While both events originated in the United States, their global relevance is evident. The U.S. economy plays a central role in international trade, financial markets, and global capital allocation. Consequently, shifts in U.S. political and trade policy can affect other G7 economies through multiple channels, including export exposure, multinational corporate earnings, exchange rate adjustments, and portfolio rebalancing. However, the magnitude and nature of these effects are likely to be heterogeneous, reflecting differences in economic structure, financial integration, and sectoral composition of stock indices.

Methodologically, this paper employs a Dynamic Conditional Correlation GARCH (DCC-GARCH) model. This approach captures changes in both conditional volatility and time-varying cross-market correlations between the U.S. market and each of the remaining G7 leading stock indices. Specifically, we estimate six USA-G7 country pairs and analyse the two exogenous shocks separately.

The main contribution of this study is threefold. First, it introduces a comparative shock-based perspective by distinguishing between two conceptually different U.S.-centred exogenous shocks, i.e., an uncertainty shock, represented by the 2024 U.S. presidential election, and an action shock, represented by the 2 April 2025 reciprocal tariff announcement. By separating a shock that primarily alters expectations and perceived uncertainty from a shock that reflects a realised policy intervention with immediate economic relevance, this paper does not treat political and policy-related events as a homogeneous class of disturbances. Second, using a DCC-GARCH model, this study measures how each shock changes conditional volatility and time-varying cross-market correlations between the United States and the other G7 stock markets. Third, by estimating six USA-G7 pairs and analysing both shocks separately, it documents cross-country differences in dynamic dependence and provides new, event-specific evidence on how different exogenous shocks shape international equity market synchronisation that is relevant for cross-border risk assessment and diversification.

This paper is organised as follows: Section 2 presents the literature review, followed by Section 3—methodology, then Section 4—results and discussion, and finally Section 5—our conclusions.

2. Literature Review

Exogenous shocks are widely treated as drivers of financial market dynamics because they reshape expectations about future cash flows, discount rates and risk premia (Antonakakis et al. 2013). In globally integrated systems, such shocks are rarely contained within national borders, but spread through trade linkages, financial exposures and shifts in risk sentiment, with contagion and spillover mechanisms documented across crises and normal times (Baig and Goldfajn 1999; Bekaert et al. 2014; Diebold and Yilmaz 2009). Political shocks form an important subset of these disturbances, encompassing elections and policy changes, and their relevance has been linked to globalisation, geopolitical tensions and populist movements (Gordell and Volgy 2022; Chan 2025).

To model the transmission of political and trade events to equity markets, it is important to distinguish between shocks that primarily raise uncertainty and those that reflect a realised policy action. Uncertainty shocks are commonly defined as discrete increases in the dispersion of expectations about future macroeconomic or policy conditions, consistent with Knightian uncertainty, in which probabilities are not precisely known (Jurado et al.

https://doi.org/10.3390/risks14040074

2015; Bloom 2009). In empirical work, uncertainty is often proxied by broad indicators such as economic policy uncertainty (Baker et al. 2016), and structural approaches separate uncertainty-driven disturbances from other macro-financial shocks (Dery and Serletis 2023; Basu and Bundick 2017). By contrast, action shocks refer to discrete and observable decisions that constitute an immediate policy intervention with direct price relevance. Relative to uncertainty shocks, their effects are typically more readily identifiable, because the decision itself is clear and the likely direction of its impact can be assessed more promptly, even if the magnitude of second-round effects remains uncertain (Kuttner 2001; Rigobon and Sack 2004; Basu and Bundick 2017; Bauer et al. 2022).

In this study, we treat the U.S. presidential election held on 5 November 2024 as an uncertainty shock, and we treat the 2 April 2025 announcement of "reciprocal tariffs" as an action shock.

Uncertainty shocks are typically defined as sudden increases in uncertainty about future macroeconomic and policy conditions, which raise perceived risk in financial markets (Jurado et al. 2015; Bloom 2009). In equity markets, this channel is commonly linked to lower valuations and higher required returns, because policy-related uncertainty can generate a systematic risk premium that investors cannot easily diversify away (Pástor and Veronesi 2012, 2013). At the same time, uncertainty shocks are often reflected in higher implied and realised volatility, and their effects can be amplified when financial frictions weaken balance sheets and tighten overall financial conditions (Jurado et al. 2015; Arellano et al. 2019; Lin et al. 2025). Elections are a good example of such a shock because they can either resolve uncertainty or intensify it, and international evidence shows that election timing is associated with changes in returns and volatility (Pantzalis et al. 2000; Białkowski et al. 2008). In the United States, presidential elections have also been linked to movements in implied volatility consistent with election-related political uncertainty (Goodell and Vähämaa 2013). Recent evidence directly treats the 2024 U.S. presidential election cycle as a prominent uncertainty episode and shows that the pre-election period is priced as an uncertainty shock (Flynn and Tarkom 2025).

Action shocks are usually understood as discrete policy actions or decisions that are observable and can trigger an immediate market response because they directly change economic conditions relevant for pricing (Kuttner 2001; Rigobon and Sack 2004). This is consistent with research that treats policy shocks as realised interventions and separates them from uncertainty about future states or policy paths (Basu and Bundick 2017; Auerbach et al. 2024; Dery and Serletis 2023). Empirical studies show that such decisions can lead to rapid asset revaluation by altering expected payoffs, especially when the decision arrives as a surprise and stock markets adjust in narrow announcement windows (Bauer et al. 2022; Klick 2025; Kuttner 2001; Rigobon and Sack 2004). Trade policy decisions can be treated as an action shock because tariffs change effective costs and market access, which matters for firms with international exposure and can induce portfolio rebalancing. Evidence from the Trump trade war documents measurable effects on U.S. financial markets (Chen et al. 2023), and related studies highlight the role of trade policy and value chains in shaping market outcomes (Blanchard et al. 2026). The trade policy literature also explicitly separates uncertainty from decision or news components, supporting the idea that realised trade actions should be analysed separately from trade policy uncertainty (Caldara et al. 2020; Graziano et al. 2024).

In the literature, the impact of shocks on equity markets has been examined from several perspectives, ranging from effects on price levels and return dynamics to price volatility and conditional variance, and from changes in cross-market dependence, measured by correlations in returns and in conditional volatilities. Some studies focus on price-level responses, reporting that political and policy-related shocks affect equity valuations and

https://doi.org/10.3390/risks14040074

generate time-varying risk premia (Pástor and Veronesi 2012, 2013; Brogaard and Detzel 2015; Pantzalis et al. 2000; Chiang 2019). Other studies shift attention from average returns to the structure of price interdependence, showing that shocks are associated with stronger co-movements across markets, particularly in stress regimes (Longin and Solnik 2001; Forbes and Rigobon 2002). Other research investigates volatility responses, consistently finding that both uncertainty-related and policy-driven shocks lead to increases in return volatility and persistent volatility clustering, producing sharp, though heterogeneous, spikes in realised or implied volatility (Bloom 2009; Ederington and Lee 1993; Białkowski et al. 2008; Goodell and Vähämaa 2013; Liu and Zhang 2015). Other studies separate realised volatility from conditional volatility, for which ARCH/GARCH models are applied (Engle 1982; Bollerslev 1986; Engle and Ng 1993; Glosten et al. 1993).

Moreover, the analysis of shock transmission across financial markets increasingly relies on multivariate GARCH models with time-varying conditional correlations, and DCC-type specifications have become standard tools for studying how dependence structures evolve during periods of heightened uncertainty. Engle (2002) introduces the Dynamic Conditional Correlation (DCC) model as an approach that combines univariate GARCH dynamics for conditional variances with a flexible parametric structure for correlations. He shows that correlations vary substantially over time and tend to increase during turbulent market conditions. Tse and Tsui (2002) propose a closely related multivariate GARCH model with autoregressive dynamics in the conditional correlation matrix, emphasising its ability to capture changes in cross-market dependence while preserving positive definiteness. Bauwens et al. (2006), in their survey of multivariate ARCH and GARCH models, stress that allowing for time-varying correlations is essential for modelling volatility spillovers and co-movements generated by common shocks across markets. Applications link time-varying correlations to uncertainty and implied volatility (Antonakakis et al. 2013), document correlation increases during global turmoil (Bekaert et al. 2014; Akhtaruzzaman et al. 2021), and show that geopolitical events can affect dynamic conditional correlations in G7-related settings (Lakhal and Zorgati 2025). Given the documented international transmission of U.S. shocks to G7 equity markets (Hanisch and Kempa 2017), we employ DCC-GARCH models to examine time-varying conditional volatility correlations across major stock indices around two distinct shocks. In particular, we compare correlation dynamics under an uncertainty shock and under an action shock. The literature supports DCC-type models as suitable tools for analysing how shocks influence the joint behaviour of market volatilities, while also highlighting remaining challenges in identifying and interpreting correlation dynamics across different shock types.

The existing literature leaves a specific research gap. Most studies focus on a single class of shocks, such as general policy uncertainty, crisis periods, geopolitical risk, or trade policy uncertainty. Others examine returns and volatility around elections or tariff announcements without directly comparing how conceptually different shocks change conditional cross-market synchronisation within the same group of countries. In the literature, there is a lack of empirical studies on whether an uncertainty shock and a realised policy action shock generate systematically different correlation responses across the same developed markets (G7 countries). Our study contributes by analysing two distinct U.S.-centred shocks and by comparing their effects on the time-varying correlations between the United States and the other G7 stock markets.

Based on the reviewed literature, two hypotheses are proposed.

H1. *The 2024 U.S. presidential election, treated as an uncertainty shock, is expected to reduce USA-G7 conditional correlations.*

https://doi.org/10.3390/risks14040074

H2. *The 2 April 2025 reciprocal tariff announcement, treated as an action shock, is expected to increase USA-G7 conditional correlations.*

3. Materials and Methods

The objective of this study is to examine how major U.S.-related shocks affect the time-varying conditional correlations between the U.S. equity market and other G7 stock markets. In particular, the analysis focuses on two distinct types of shocks: an uncertainty shock, represented by the U.S. presidential election in November 2024, and an action shock, represented by the announcement of U.S. tariff increases in April 2025. While the former primarily reflects political uncertainty and expectations, the latter constitutes a direct economic policy intervention with potentially global spillover effects.

The empirical analysis is conducted using daily stock market indices for the United States and six other G7 economies, i.e., Canada, France, Germany, Italy, Japan, and the United Kingdom. The U.S. market is represented by the S&P 500, while the international markets include TSX 60 (Canada), FTSE All Share (United Kingdom), DAX (Germany), CAC All (France), FTSE Italia All Share (Italy), and TOPIX (Japan). The sample covers the period from January 2010 to July 2025 and consists of daily observations.

All price series are transformed into logarithmic returns, computed as the first difference of the natural logarithm of index prices. This transformation ensures stationarity and allows for a consistent interpretation of returns across markets.

This study employs the Dynamic Conditional Correlation GARCH model (DCC-GARCH) (Engle 2002) to analyse time-varying dependence between the United States and other G7 stock market leading indices. The model allows for jointly modelling conditional variances and dynamic correlations of financial return series.

Let y_t denote a vector of asset logarithmic returns. The conditional mean and variance structure is given by

$$y_t = \mu_t + \varepsilon_t, \varepsilon_t \sim N(0, \in_t) \tag{1}$$

$$\in_t = D_t P_t D_t \tag{2}$$

where D_t is a diagonal matrix of conditional standard deviations ($D_t = diag(\sqrt{h_{1t}}, \sqrt{h_{2t}}, \ldots, \sqrt{h_{Nt}},)$ and P_t is the time-varying correlation matrix. Each conditional variance follows a univariate GARCH(1,1) process:

$$h_{1t} = \gamma_i + \alpha_i \varepsilon_{i,t-1}^2 + \beta_i h_{i,t-1};\ i = 1, 2, \ldots, N \tag{3}$$

Dynamic conditional correlations are obtained from standardised residuals and summarised by the bivariate correlation coefficient ρ_t, which captures the time-varying dependence between asset returns.

Following Engle (2002), the dynamics of the conditional correlation matrix are modelled using the DCC(1,1) specification:

$$Q_t = (1 - a - b)\overline{Q} + a z_{t-1} z_{t-1}^T + b Q_{t-1} \tag{4}$$

where $z_t = D_t^{-1} \varepsilon_t$ denotes the vector of standardized residuals, $\overline{Q}$ is the unconditional covariance correlation matrix of z_t, and Q_t is the intermediate time-varying covariance matrix.

The dynamic conditional correlation matrix P_t is obtained by rescaling Q_t as

$$P_t = diag(Q_t)^{-1/2} Q_t diag(Q_t)^{-1/2} \tag{5}$$

The parameters a and b govern the dynamics of conditional correlation, where a measures the impact of new shocks on correlation. While b captures the persistence of

https://doi.org/10.3390/risks14040074

correlation dynamics. The condition $a + b < 1$ ensures mean reversion and stationarity of the dynamic correlation.

The DCC-GARCH model is estimated over the full sample period (2010–2025) to capture long-run volatility and the dynamics of correlation between the USA and each G7 market. The study period starts in January 2010 to ensure a sufficiently long post-global financial crisis period for stable estimation of GARCH and DCC parameters across all U.S.-G7 bilateral pairs, while ending in 2025 so as to include the post-event adjustment following the April 2025 reciprocal tariff announcement. Based on the estimated model, time-varying conditional correlations are extracted for each country–USA pair.

To provide a clearer visualisation of correlation dynamics around the selected shocks, dynamic conditional correlation paths are plotted for the period July 2024 to July 2025, which encompasses both events of interest, i.e., the 2024 U.S. presidential election and President Trump's announcement of "reciprocal tariffs". The selected time window enables more precise identification of changes in correlation patterns associated with the U.S. presidential election and the subsequent tariff announcement. In addition to visual inspection, an event study approach is used to measure changes in conditional correlations. For each event, a symmetric $\pm$30-day (calendar days) window around the event date is considered. The window length was selected as a compromise between capturing immediate event-related adjustments and limiting contamination from unrelated or random developments. Average conditional correlations are computed separately for the pre-event and post-event subperiods.

A positive change in average conditional correlation indicates an increase in market co-movement, suggesting stronger financial integration or a synchronised response to the shock. Conversely, a negative change implies a reduction in co-movement, which may be interpreted as market decoupling or increased idiosyncratic uncertainty. The magnitude of the correlation change reflects the economic relevance of the shock, with larger absolute values indicating a stronger impact on cross-market dependence.

To complement the descriptive comparison of pre-event and post-event average conditional correlations, we additionally test whether the observed changes are statistically significant. For each country/USA pair and each event window, we estimate a simple regression as

$$\rho_t = \alpha + \beta D_t + \varepsilon_t \tag{6}$$

where ρ_t denotes the dynamic conditional correlation obtained from the DCC-GARCH model and D_t is a dummy variable equal to 1 for observations after the event date and 0 otherwise. The coefficient β captures the post-event change in the mean conditional correlation. Since the event window series may exhibit heteroskedasticity and serial dependence, statistical inference is based on heteroskedasticity-consistent and autocorrelation-consistent (HAC) standard errors computed using the Newey–West estimator.

As an additional robustness check, we also apply the Welch two-sample t-test to compare pre-event and post-event mean conditional correlations. This test does not assume equal variances across the two subperiods and serves as a complementary benchmark for the statistical significance of the event-related changes.

By combining full-sample DCC-GARCH estimation with event-centred analysis, this approach allows for assessing both the persistent nature of correlation dynamics and the short-term impact of specific uncertainty and action shocks on USA-G7 stock market interdependence.

4. Results and Discussion

To examine the dynamics of conditional volatility and time-varying dependence between the U.S. stock market and other G7 equity markets, Dynamic Conditional Correlation

https://doi.org/10.3390/risks14040074

GARCH models are estimated for pairs consisting of a U.S. stock index and one selected G7 stock index, yielding six bilateral market pairs. As a prerequisite for the DCC model, univariate volatility models are first estimated for each return series. Table 1 reports the estimated parameters of the univariate GARCH(1,1) variance equations defined in Equation (3) for the equity indices. Table 1 presents the estimates of the variance constant γ_i, the ARCH parameter α_i capturing the impact of past shocks, and the GARCH parameter β_i measuring volatility persistence through lagged conditional variance. The sum $\alpha_i + \beta_i$ is also reported to assess the degree of volatility persistence implied by the model.

Table 1. Estimated GARCH(1,1) and DCC parameters for U.S.-G7 pairs.

Country	Index	γ_i	α_i	β_i	α_i+β_i
Canada	TSX 60	0.000002 * (0.000001)	0.1386 *** (0.0222)	0.8312 *** (0.0239)	0.9698
UK	FTSE All Share	0.000004 *** (0.000001)	0.1440 *** (0.0122)	0.8032 *** (0.0206)	0.9472
Germany	DAX	0.000004 (0.000108)	0.1051 (0.1201)	0.8666 (0.8055)	0.9717
France	CAC All	0.000005 (0.000005)	0.1487 *** (0.0184)	0.8174 *** (0.0364)	0.9661
Italy	FTSE Italia All Share	0.000005 (0.000004)	0.1145 *** (0.0127)	0.8584 *** (0.0281)	0.9729
Japan	TOPIX	0.000008 *** (0.000001)	0.1351 *** (0.0108)	0.8060 *** (0.0136)	0.9411

Note: Standard errors in parentheses. ***, **, * denote significance at the 1%, 5%, and 10% levels, respectively. Source: Own calculation based on data from Refinitiv Eikon (LSEG).

The estimated GARCH(1,1) parameters reported in Table 1 indicate a high degree of volatility persistence across all G7 equity markets. In all cases, the ARCH coefficients (α) are positive and moderate in size, suggesting that new shocks to returns have a noticeable but limited short-run impact on conditional volatility. The GARCH coefficients (β) are consistently large, implying that volatility responds strongly to its own past realisations. As a result, the sum $\alpha + \beta$ is close to unity for all markets, indicating slow mean reversion and long-lasting volatility effects. For most non-U.S. markets, the ARCH and GARCH coefficients are statistically significant, confirming that past shocks and lagged conditional variance play an important role in explaining volatility dynamics. The evidence is particularly strong for Canada, the UK, Italy, and Japan, whereas the German specification is less precisely estimated. The results confirm that the GARCH(1,1) specification provides an adequate representation of conditional variance dynamics for all examined stock indices.

Based on the standardised residuals obtained from estimated GARCH(1,1) models, in the second step, we estimate the Dynamic Conditional Correlation structure. Table 2 presents the estimated DCC parameters, as defined in Equation (4), for each USA-G7 equity index pair. Table 2 reports the shock parameter a, which measures the impact of new shocks on correlation, while b captures the persistence of correlation dynamics. The sum $a + b$ is reported to evaluate the persistence and stationarity of the dynamic conditional correlation process.

The DCC parameter estimates reported in Table 2 reveal highly persistent correlation dynamics between the USA and other G7 equity markets. The shock parameter a is relatively small across all country pairs and is significant only for selected pairs, indicating that conditional correlations respond only modestly to new information or short-term shocks. In contrast, the persistence parameter b is positive, close to unity, and highly

https://doi.org/10.3390/risks14040074

statistically significant in all cases, leading to values of $a + b$ that approach one while remaining below the stationarity threshold. This implies that changes in cross-market correlations are highly persistent and tend to vanish slowly over time. The strongest persistence is observed for the USA-UK and USA/Germany pairs, while relatively lower (but still substantial) persistence characterises the USA/Canada correlation. These findings suggest that correlation dynamics among developed equity markets are dominated by long-run integration effects rather than transitory shocks, making them particularly susceptible to prolonged periods of heightened co-movement following major global events.

Table 2. DCC correlation parameters for country/USA pairs in G7 countries.

Country/USA Pair	*a*	*b*	*a+b*
Canada/USA	0.0325 *** (0.0089)	0.9037 *** (0.0266)	0.9362
UK/USA	0.0090 (0.0093)	0.9860 *** (0.0177)	0.9950
Germany/USA	0.0085 (0.0057)	0.9869 *** (0.0115)	0.9954
France/USA	0.0204 * (0.0117)	0.9622 *** (0.0300)	0.9826
Italy/USA	0.0164 *** (0.0055)	0.9669 *** (0.0177)	0.9833
Japan/USA	0.0045 (0.0048)	0.9522 *** (0.0300)	0.9567

Note: ***, **, * denote significance at the 1%, 5%, and 10% levels, respectively. Source: Own calculation based on data from Refinitiv Eikon (LSEG).

The DCC-GARCH model was estimated using the full sample of daily data from 1 January 2010 to 31 December 2025, allowing for a characterisation of long-run volatility and correlation dynamics between the U.S. equity market and other G7 markets. Based on these estimates, time-varying dynamic conditional correlations are obtained for each country/USA pair.

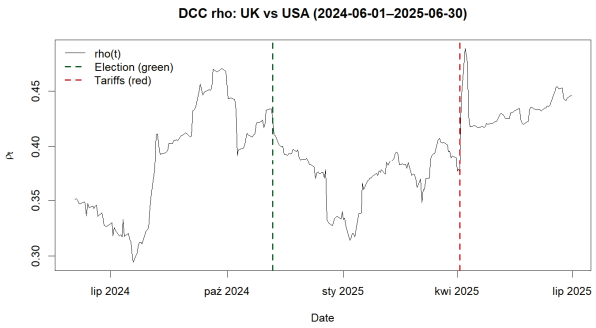
In the next step, we aim to verify the effect of two analysed shocks (uncertainty shock and action shock) on dynamic conditional correlation for each country/USA pair. To better capture the effects of these shocks, the analysis of correlation dynamics is presented graphically for a restricted time window from 1 June 2024 to 30 June 2025. This window encompasses the two events of interest and allows for a clearer visualisation of changes in conditional correlations around their occurrence. The dynamic conditional correlation plots include vertical reference lines marking the timing of the analysed shocks. The U.S. presidential election held on 5 November 2024 is indicated by a green dashed vertical line and represents a 2024 presidential election uncertainty shock. President Trump's announcement of U.S. tariff increases on 2 April 2025 is marked by a red dashed vertical line and captures a trade policy action shock with potentially broad international spillover effects (Figure 1A–F).

Figure 1 reveals that baseline dependence differs markedly across pairs, with Canada/USA showing the highest and most persistent correlation throughout the window, Japan/USA remaining consistently the lowest, and the European pairs lying in between. Second, the two analysed shocks are associated with clearly different correlation responses. Around the 2024 U.S. presidential election date, the dominant pattern is a decline in conditional correlation for the European markets, indicating temporary decoupling and more idiosyncratic dynamics during the uncertainty event period, with the largest and

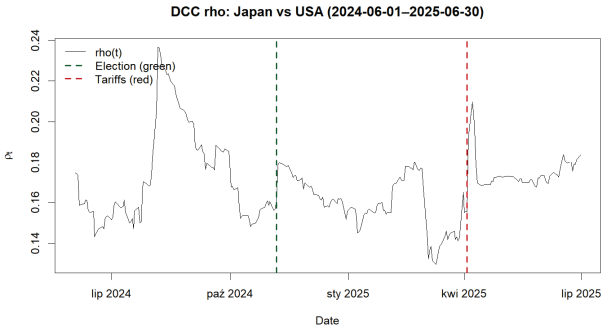
https://doi.org/10.3390/risks14040074

most persistent post-election drops visible for France/USA and Italy/USA, a pronounced but less extreme decline for Germany/USA, and a milder downward adjustment for the UK-USA. By contrast, Canada/USA and Japan/USA react only modestly around the election, suggesting limited changes in co-movement for these two pairs in response to political uncertainty.

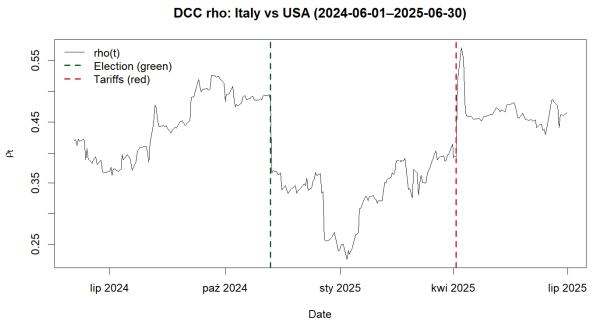
(**A**) UK/USA



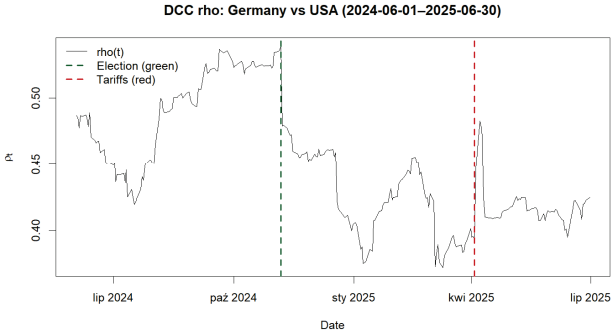
(**B**) Japan/USA



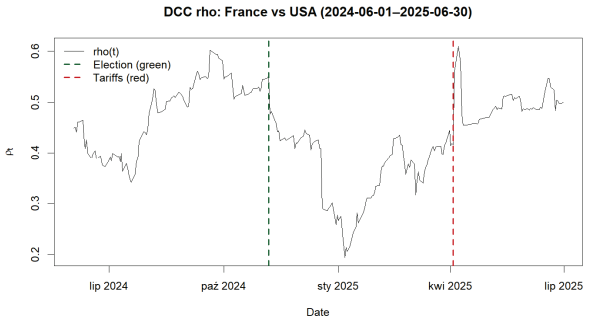
(**C**) Italy/USA



(**D**) Germany/USA



(**E**) France/USA



(**F**) Canada/USA

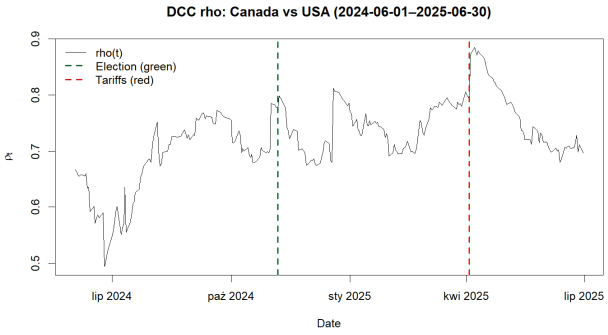


Figure 1. DCC-GARCH dynamic conditional correlations for USA-G7 equity index pairs around the 2024 U.S. presidential election and the 2025 reciprocal tariffs announcement. Source: Own calculation and elaboration based on data from Refinitiv Eikon (LSEG).

Around President Trump's tariff announcement on 2 April 2025, the pattern reverses across all six country/USA pairs, with an immediate upward jump in conditional correlations that is visible for every analysed pair, consistent with a common, policy action synchronising markets. The magnitude of this action shock response is clearly heterogeneous, i.e., it is strongest for Canada/USA (a sharp move to a local peak followed by gradual mean reversion), very large for France/USA and Italy/USA (a steep increase and a higher post-event level than in the months preceding the announcement), moderate for the UK-USA and Germany/USA, and smallest for Japan–USA, where the increase is noticeable but remains limited in absolute terms. In summary, the election-related uncertainty shock

is associated mainly with a weakening of correlations concentrated in Europe, whereas the tariff-related action shock generates a broad, positive correlation shift across all country/USA pairs, with the largest synchronisation effects in North America and parts of Europe, and the weakest in Japan.

In the next step, we compute average conditional correlations over a symmetric ±30-day (calendar days) window around each of the two analysed event dates using DCC-GARCH estimates (Table 3). In addition to reporting pre-event and post-event mean correlations, Table 3 includes t-statistics based on HAC (Newey–West) standard errors and Welch two-sample t-statistics, both accompanied by significance indicators.

Table 3. Average conditional correlations in the ±30-day event window.

Country/USA Pair	Event	Pre-Event Mean ρ	Post-Event Mean ρ	Change (Post-Event–Pre-Event)	t HAC Statatistic	Welch Test Statistic
Canada/USA	2024 election	0.711	0.721	0.010	0.584	0.374
France/USA	2024 election	0.526	0.438	−0.088	−12.840 ***	1.436 ***
Germany/USA	2024 election	0.525	0.462	−0.063	−18.903 ***	6.156 ***
Italy/USA	2024 election	0.488	0.348	−0.140	−30.056 ***	1.250 ***
Japan/USA	2024 election	0.155	0.172	0.017	6.998 ***	9.795 ***
UK-USA	2024 election	0.416	0.394	−0.022	−4.476 ***	3.193 ***
Canada/USA	2025 tariffs	0.785	0.838	0.053	4.055 ***	2.731 ***
France/USA	2025 tariffs	0.390	0.486	0.096	4.538 ***	1.342 ***
Germany/USA	2025 tariffs	0.387	0.424	0.037	4.140 ***	7.226 ***
Italy/USA	2025 tariffs	0.381	0.477	0.096	6.161 ***	1.870 ***
Japan/USA	2025 tariffs	0.142	0.177	0.034	6.933 ***	1.434 ***
UK-USA	2025 tariffs	0.387	0.430	0.043	4.167 ***	2.827 ***

Note: ***, **, * denote significance at the 1%, 5%, and 10% levels, respectively. Source: Own calculation based on data from Refinitiv Eikon (LSEG).

The results in Table 3 for the 2024 U.S. presidential election indicate a heterogeneous but statistically significant response across markets. For the European pairs, i.e., France–USA, Germany–USA, Italy–USA, and UK–USA, the post-event changes in conditional correlations are negative and statistically significant under both the HAC and Welch tests. The strongest decline is observed for Italy–USA, followed by France–USA and Germany–USA. In contrast, the change for Canada–USA is small and not statistically significant, whereas Japan–USA shows a modest, statistically significant increase in correlation. A markedly different pattern emerges for the 2025 tariff announcement. In all country–USA pairs, post-event mean correlations exceed their pre-event values, and the increases are statistically significant across both HAC and Welch tests. The largest increases are observed for France–USA and Italy–USA, followed by Canada–USA and the UK–USA, while Japan–USA shows the smallest, yet still significant, increase. In general, the results in Table 3 support the interpretation that the election uncertainty shock leads to heterogeneous, region-specific correlation responses, while the trade policy action shock acts as a common disturbance that significantly strengthens international market synchronisation across the G7 markets.

These differences may reflect cross-country variation in economic and financial linkages with the United States, including differences in trade exposure, market integration, and index composition. Canada's particularly strong response to the tariff announcement

https://doi.org/10.3390/risks14040074

is consistent with its exceptionally high degree of North American trade and financial integration, which makes its equity market especially sensitive to realised U.S. policy actions. The comparatively strong increases observed for France and Italy may indicate greater sensitivity to global growth repricing and external demand channels, while the more moderate responses of Germany and the UK may be associated with differences in sectoral composition, multinational diversification, and the relative importance of domestic market factors. Japan's weaker response is economically plausible given its lower direct exposure to U.S.-centred trade policy transmission, its distinct regional market environment, and the stronger role of domestic and Asia-specific determinants of equity valuation. The election-related uncertainty shock appears to have operated differently, producing weaker correlations mainly in Europe, where investors may have revised U.S.-related risks in a more heterogeneous and less synchronised way.

Our findings align with prior evidence that election periods can be associated with shifts in co-movement and volatility, and that such effects are heterogeneous across countries (Pantzalis et al. 2000; Białkowski et al. 2008; Goodell and Vähämaa 2013). In our results, the 2024 election-related uncertainty shock is mainly associated with lower conditional correlations for the European pairs, indicating a temporary decline in synchronisation with the U.S. market. This is consistent with the view that international dependence is state-dependent and can weaken around shocks that raise uncertainty and generate more idiosyncratic dynamics (Longin and Solnik 2001; Forbes and Rigobon 2002). By contrast, the tariff announcement is followed by increases in correlation across all pairs, consistent with evidence that trade policy settings distinguish realised decision/news components from uncertainty components and that such decisions can operate as common shocks (Caldara et al. 2020).

From a hypothesis-testing perspective, the study findings provide partial support for H1 and strong support for H2. The analysed election-related uncertainty shock reduces correlations only in part of the sample, primarily among the European G7 markets, while the tariff-related action shock produces a uniform and statistically significant increase in conditional correlations across all U.S.-G7 pairs. The results indicate that uncertainty shocks and action shocks differ not only in magnitude but also in the cross-sectional consistency of their effects on international equity market synchronisation.

5. Conclusions

Stock markets respond to major exogenous shocks, which can alter cross-market co-movement. In this study, we distinguish two exogenous shock types from the perspective of financial markets. The uncertainty shock is linked to political uncertainty around the U.S. presidential election in November 2024. The action shock is linked to a discrete trade policy intervention, namely President Trump's announcement of reciprocal tariffs in April 2025.

This study examines how these shocks affect time-varying conditional correlations between the U.S. equity market and other G7 markets. We use daily data for the S&P 500 and leading equity indices for Canada, France, Germany, Italy, Japan and the United Kingdom from January 2010 to July 2025. Conditional correlation is modelled with DCC-GARCH and estimated for six USA-G7 pairs, and shock effects are assessed with correlation paths for 1 June 2024 to 30 June 2025 and a symmetric $\pm$30-day event window.

The results confirm that both analysed geopolitical events significantly affect the dynamic correlations between the USA and other G7 equity markets. The event analysis indicates that the election-related uncertainty shock mainly reduces conditional correlations for European pairs, with the largest decline for Italy and clear drops for France and Germany, consistent with temporary decoupling and more idiosyncratic dynamics. Canada

https://doi.org/10.3390/risks14040074

and Japan show only small changes around the election. By contrast, the tariff-related action shock significantly increases conditional correlations for every analysed country/USA pair, with the largest rises for France and Italy, a sizeable increase for Canada, and the smallest adjustment for Japan. While the 2024 U.S. presidential election is associated with a decline in correlations for European markets, the 2025 tariff announcement leads to a widespread increase in cross-market synchronisation. These effects are statistically significant under the HAC (Newey–West) and Welch tests. This highlights that different types of shocks generate distinct patterns of international market co-movement. Stock market synchronisation weakens in response to the uncertainty shock and strengthens in response to the action shock.

Our findings have some implications for international portfolio allocation and financial risk management. Higher conditional correlations observed after action shocks suggest that realised policy interventions may compress diversification benefits more strongly than uncertainty-driven events, as they induce more synchronous cross-market repricing and strengthen spillover effects across major equity markets. Consequently, investors and risk managers should not treat political and policy-related events as a homogeneous category of shocks. Instead, scenario analysis and portfolio stress testing should explicitly distinguish between uncertainty-driven episodes and realised policy actions, since the latter may generate more immediate and broader increases in market interdependence.

This study has some limitations. Firstly, it focuses on six bilateral G7 pairs and two specific U.S.-centred events, which limits generalisability. Moreover, event window evidence cannot fully isolate concurrent issues, including, e.g., macro-financial news. Additionally, broad market indices may conceal sector-level heterogeneity in shock transmission.

This paper's contribution is to classify the two events as different shock types and to document opposite correlation responses across the same set of developed markets within a DCC-GARCH approach. To the best of our knowledge, such a comparison for these events within the G7 setting has not been provided before. Future research can extend the analysis to additional countries and shocks and assess robustness using alternative correlation models, including asymmetric DCC-GARCH specifications.

Author Contributions: Conceptualisation, K.C. and M.W.; methodology, K.C. and M.W.; software, K.C. and M.W.; validation, K.C. and M.W.; formal analysis, K.C.; investigation, K.C. and M.W.; resources, K.C. and M.W.; data curation, K.C. and M.W.; writing—original draft preparation, K.C. and M.W.; writing—review and editing, K.C. and M.W.; visualisation, K.C. and M.W.; supervision, K.C. and M.W.; project administration, K.C. and M.W.; funding acquisition, K.C. and M.W. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The data analysed in this study were obtained from Refinitiv Eikon (LSEG) under a paid subscription and are not publicly available. Access to the data can be obtained directly from LSEG/Refinitiv, subject to the provider's terms and conditions.

Conflicts of Interest: The authors declare no conflicts of interest.

References

Agoraki, Maria-Eleni. K., Georgios P. Kouretas, and Nikiforos T. Laopodis. 2022. Geopolitical risks, uncertainty, and stock market performance. *Economic and Political Studies* 10: 253–65. [CrossRef]

Akhtaruzzaman, Md, Sabri Boubaker, and Ahmet Sensoy. 2021. Financial contagion during COVID–19 crisis. *Finance Research Letters* 38: 101604. [CrossRef]

Antonakakis, Nikolaos, Ioannis Chatziantoniou, and George Filis. 2013. Dynamic co-movements of stock market returns, implied volatility and policy uncertainty. *Economics Letters* 120: 87–92. [CrossRef]

Arellano, Cristina, Yan Bai, and Patrick J. Kehoe. 2019. Financial frictions and fluctuations in volatility. *Journal of Political Economy* 127: 2049–103. [CrossRef]
Arouri, Mohamed, Christophe Estay, Christophe Rault, and David Roubaud. 2016. Economic policy uncertainty and stock markets: Long-run evidence from the US. *Finance Research Letters* 18: 136–41. [CrossRef]
Auerbach, Alan J., Yuriy Gorodnichenko, and Daniel Murphy. 2024. Macroeconomic frameworks: Reconciling evidence and model predictions from demand shocks. *American Economic Journal: Macroeconomics* 16: 190–29. [CrossRef]
Baig, Taimur, and Ilan Goldfajn. 1999. Financial market contagion in the Asian crisis. *IMF Staff Papers* 46: 167–95. [CrossRef]
Baker, Scott R., Nicholas Bloom, and Steven J. Davis. 2016. Measuring economic policy uncertainty. *The Quarterly Journal of Economics* 131: 1593–636. [CrossRef]
Basu, Susanto, and Brent Bundick. 2017. Uncertainty shocks in a model of effective demand. *Econometrica* 85: 937–58. [CrossRef]
Bauer, Michael D., Aeimit Lakdawala, and Philippe Mueller. 2022. Market-based monetary policy uncertainty. *The Economic Journal* 132: 1290–308. [CrossRef]
Bauwens, Luc, Sébastien Laurent, and Jeroen V. Rombouts. 2006. Multivariate GARCH models: A survey. *Journal of Applied Econometrics* 21: 79–109. [CrossRef]
Bekaert, Geert, Michael Ehrmann, Marcel Fratzscher, and Arnaud Mehl. 2014. The global crisis and equity market contagion. *The Journal of Finance* 69: 2597–649. [CrossRef]
Białkowski, Jędrzej, Katrin Gottschalk, and Tomasz P. Wisniewski. 2008. Stock market volatility around national elections. *Journal of Banking & Finance* 32: 1941–53. [CrossRef]
Blanchard, Emily J., Chad P. Bown, and Robert C. Johnson. 2026. Global value chains and trade policy. *Review of Economic Studies* 93: 181–214. [CrossRef]
Bloom, Nicholas. 2009. The impact of uncertainty shocks. *Econometrica* 77: 623–85. [CrossRef]
Bollerslev, Tim. 1986. Generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics* 31: 307–27. [CrossRef]
Brogaard, Jonathan, and Andrew Detzel. 2015. The asset-pricing implications of government economic policy uncertainty. *Management Science* 61: 3–18. [CrossRef]
Caldara, Dario, Matteo Iacoviello, Patrick Molligo, Andrea Prestipino, and Andrea Raffo. 2020. The economic effects of trade policy uncertainty. *Journal of Monetary Economics* 109: 38–59. [CrossRef]
Chan, Ka M. 2025. A tale of two cities: Spillover effects of electoral shocks in non-democratic regimes. *Democratization* 32: 27–52. [CrossRef]
Chen, Yong, Jing Fang, and Dingming Liu. 2023. The effects of Trump's trade war on US financial markets. *Journal of International Money and Finance* 134: 102842. [CrossRef]
Chiang, Thomas C. 2019. Economic policy uncertainty, risk and stock returns: Evidence from G7 stock markets. *Finance Research Letters* 29: 41–49. [CrossRef]
Cunha, Raphael, and Andreas Kern. 2022. Global banking and the spillovers from political shocks at the core of the world economy. *The Review of International Organizations* 17: 717–49. [CrossRef]
Dery, Cosmas, and Apostolos Serletis. 2023. Macroeconomic fluctuations in the United States: The role of monetary and fiscal policy shocks. *Open Economies Review* 34: 961–77. [CrossRef]
Diebold, Francis X., and Kamil Yilmaz. 2009. Measuring financial asset return and volatility spillovers, with application to global equity markets. *The Economic Journal* 119: 158–71. [CrossRef]
Dungey, Mardi, Faisal Khan, and Mala Raghavan. 2018. International trade and the transmission of shocks: The case of ASEAN-4 and NIE-4 economies. *Economic Modelling* 72: 109–21. [CrossRef]
Ederington, Louis H., and Jae H. Lee. 1993. How markets process information: News releases and volatility. *The Journal of Finance* 48: 1161–91. [CrossRef]
Engle, Robert F. 1982. Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation. *Econometrica* 50: 987–1007. [CrossRef]
Engle, Robert F. 2002. Dynamic conditional correlation: A simple class of multivariate generalized autoregressive conditional heteroskedasticity models. *Journal of Business & Economic Statistics* 20: 339–50.
Engle, Robert F., and Victor K. Ng. 1993. Measuring and testing the impact of news on volatility. *The Journal of Finance* 48: 1749–78. [CrossRef]
Flynn, Matthew, and Augustine Tarkom. 2025. How do financial markets price political uncertainty? Evidence from the 2024 United States presidential election. *Finance Research Letters* 75: 106879. [CrossRef]
Forbes, Kristin J., and Roberto Rigobon. 2002. No contagion, only interdependence: Measuring stock market comovements. *The Journal of Finance* 57: 2223–61. [CrossRef]
Glosten, Lawrence R., Ravi Jagannathan, and David E. Runkle. 1993. On the relation between the expected value and the volatility of the nominal excess return on stocks. *The Journal of Finance* 48: 1779–801. [CrossRef]

Goodell, John W., and Sami Vähämaa. 2013. US presidential elections and implied volatility: The role of political uncertainty. *Journal of Banking & Finance* 37: 1108–117. [CrossRef]
Gordell, Kelly M., and Thomas J. Volgy. 2022. Political shocks in foreign policy and international politics: An alternative approach. *Canadian Foreign Policy Journal* 28: 109–26. [CrossRef]
Graziano, Alejandro. G., Kyle Handley, and Nuno Limão. 2024. An import (ant) price of Brexit uncertainty. *Journal of International Economics* 152: 104012. [CrossRef]
Hanisch, Max, and Bernd Kempa. 2017. The international transmission channels of US supply and demand shocks: Evidence from a non-stationary dynamic factor model for the G7 countries. *The North American Journal of Economics and Finance* 42: 70–88. [CrossRef]
Jurado, Kyle, Sydney C. Ludvigson, and Serena Ng. 2015. Measuring uncertainty. *American Economic Review* 105: 1177–216. [CrossRef]
Klick, Jonathan. 2025. Market response to court rejection of California's board diversity laws. *Journal of Empirical Legal Studies* 22: 4–26. [CrossRef]
Kuttner, Kenneth N. 2001. Monetary policy surprises and interest rates: Evidence from the Fed funds futures market. *Journal of Monetary Economics* 47: 523–44. [CrossRef]
Lakhal, Chiraz, and Imen Zorgati. 2025. Russia-Ukraine conflict, commodities and stock market: DCC-GARCH approach. *International Journal of Computational Economics and Econometrics* 15: 437–57. [CrossRef]
Lin, Shih-Yang, Yi-Chan Tsai, and Po-Yuan Wang. 2025. Financial frictions and uncertainty shocks. *Macroeconomic Dynamics* 29: e143. [CrossRef]
Liu, Li, and Tao Zhang. 2015. Economic policy uncertainty and stock market volatility. *Finance Research Letters* 15: 99–105. [CrossRef]
Longin, François, and Bruno Solnik. 2001. Extreme correlation of international equity markets. *The Journal of Finance* 56: 649–76. [CrossRef]
Mei, Jianping, and Limin Guo. 2004. Political uncertainty, financial crisis and market volatility. *European Financial Management* 10: 639–57. [CrossRef]
Pantzalis, Christos, David A. Stangeland, and Harry J. Turtle. 2000. Political elections and the resolution of uncertainty: The international evidence. *Journal of Banking & Finance* 24: 1575–604. [CrossRef]
Pástor, L'uboš, and Pietro Veronesi. 2012. Uncertainty about government policy and stock prices. *The Journal of Finance* 67: 1219–64. [CrossRef]
Pástor, L'uboš, and Pietro Veronesi. 2013. Political uncertainty and risk premia. *Journal of Financial Economics* 110: 520–45. [CrossRef]
Rigobon, Roberto, and Brian Sack. 2004. The impact of monetary policy on asset prices. *Journal of Monetary Economics* 51: 1553–75. [CrossRef]
Tse, Yiu K., and Albert K. C. Tsui. 2002. A multivariate generalized autoregressive conditional heteroscedasticity model with time-varying correlations. *Journal of Business & Economic Statistics* 20: 351–62. [CrossRef]

risks

MDPI

Article

Estimating Corporate Bond Market Volatility Using Asymmetric GARCH Models

Elroi Hadad [1,*], Amit Malka Fridman [2] and Rami Yosef [2]

[1] Department of Industrial Engineering and Management, Sami Shamoon College of Engineering, Beer-Sheva 8410802, Israel
[2] Department of Business Administration, Guilford Glazer Faculty of Business and Management, Ben-Gurion University of the Negev, Beer-Sheva 8410501, Israel; ramiyo@bgu.ac.il (R.Y.)
* Correspondence: hadadel@sce.ac.il

Abstract

This study investigates the volatility of the Israeli corporate bond market, where corporate bonds are traded on a Limit Order Book (LOB) exchange with high retail trading activity. Using data from the Tel-Bond 20 and Tel-Bond 60 indices, we estimate various asymmetric GARCH models to capture the dynamics of bond returns. Our findings highlight a leverage effect, where negative shocks have a more significant impact on volatility than positive shocks, underscoring the importance of investor sentiment. The GJR model with a Student's *t*-distribution best captures serial correlation, persistence of conditional volatility, and asymmetric volatility clustering. These results have significant implications for risk management, portfolio allocation, and regulatory policies, emphasizing the need for robust volatility forecasting models in transparent and active corporate bond markets.

Keywords: corporate bonds; market efficiency; investor sentiment; volatility; GARCH modeling

JEL Classification: C58; G12; G14

1. Introduction

Forecasting volatility in financial markets remains a cornerstone of financial research, driven by its critical importance to asset pricing, portfolio management, and financial stability. While the role of investor sentiment in the volatility of stock returns has been extensively studied (Brown and Cliff 2004; Hadad and Kedar-Levy 2024; Verma and Verma 2007), empirical evidence on how sentiment-driven dynamics influence corporate bond market volatility remains limited. This gap is surprising given the rapid expansion of corporate debt markets globally and their rising exposure to behavioral trading activity, particularly from retail investors (Ederington et al. 2015; Liu et al. 2022).

Recent studies show varying degrees of sentiment's impact on U.S. corporate bond yield (Bethke et al. 2017; Nayak 2010; Piñeiro-Chousa et al. 2022) and on European and Chinese yield spreads (Clayton et al. 2009; Lee 2019; Spyrou 2013), highlighting the potential impact of investors behavior on debt markets. Other research indicates that announcement shocks have a strong impact on bond market volatility dynamics, suggesting that bond investors incorporate news faster than other information (de Goeij and Marquering 2006). Intuitively, the sentiment effect in corporate bond markets affects the volatility of returns, given that these investors may be viewed as noise traders (Foucault et al. 2011; Kumar and Lee 2006; Yung and Nafar 2017). Indeed, an extant body of literature has found that changes in sentiment may be associated with changes in stock return volatility, often, but

not always, with a positive sign (Brown and Cliff 2004; Baker and Wurgler 2006; Baker and Wurgler 2007). However, the impact on corporate bond markets is rarely discussed.

While knowledge about volatility behavior in corporate bond markets is scarce, financial literature acknowledges its economic implications. Volatility in bond markets significantly impacts investors and the broader economy. For investors, increased volatility translates to higher uncertainty and risk, potentially leading to greater returns or losses (Attarzadeh and Balcilar 2022; Bai et al. 2021). High volatility can affect bond pricing, making it difficult for investors to predict future prices and yields accurately (Pham and Cepni 2022). This uncertainty necessitates sophisticated risk management strategies to hedge against potential losses. Additionally, volatility in corporate bonds impacts portfolio allocation decisions, leading investors to shift their preferences toward more stable assets during volatile periods (Dick-Nielsen et al. 2012; Friewald et al. 2012).

Economically, volatility in the bond market can influence the cost of borrowing for corporations (Turkmen Muldur et al. 2019). Higher volatility can lead to increased risk premiums, raising the cost of issuing new debt (Bai et al. 2021). This, in turn, can affect corporate investment and growth strategies. Moreover, significant fluctuations in bond prices can affect financial market stability, influencing the decisions of policymakers (Abudy and Shust 2023). Given that corporate bond trading has been on the rise, particularly after the sub-prime crisis (Choi and Kim 2018; Graham et al. 2015), it is crucial to investigate bond market volatility for efficient economic stability and investor confidence.as

In this paper, we explore if well-established volatility forecasting models in stock markets are equally applicable to corporate bonds. Intuitively, sentiment effects on corporate bond returns would be more evident in markets with high participation rates of individual, as retail-sized investors are more prone to sentiment (Baker and Wurgler 2007). The Israeli corporate bond market provides an ideal setting for analyzing volatility behavior, as it is characterized by depth, high transparency, and high retail trading activity (Abudy and Wohl 2018; Gur-Gershgoren et al. 2020). Unlike the worldwide practice where corporate bonds are traded on a separate over-the-counter (OTC) market, bonds in the Tel Aviv Stock Exchange (TASE) are traded on a limit order book (LOB) trading system like stocks. This market design increases price discovery efficiency but may also intensify intraday volatility due to frequent order updates and the limited ability of institutional investors to provide stabilizing liquidity (Abudy and Shust 2023).

Recent evidence suggests that corporate bonds traded on the TASE exhibit stronger volatility transmission and contagion effects compared to traditional OTC markets, primarily due to TASE's unique LOB mechanism. Abudy and Wohl (2018) show that corporate bonds on TASE display higher liquidity and narrower bid-ask spreads than their U.S. OTC counterparts, attributing these advantages to the high participation of retail investors and the centralized, transparent nature of the LOB structure. (Hadad and Kedar-Levy 2024) further demonstrate that changes in investor sentiment, particularly among retail participants, significantly influence conditional return volatility in both bond and stock markets, with the effects varying in magnitude and direction depending on prevailing market conditions. (Hadad 2025) provides further evidence that the LOB mechanism facilitates cross-asset contagion between stocks and bonds, especially during periods of market stress, highlighting how sentiment-driven trades in a retail-dominated environment can elevate systemic risk. These findings underscore the importance of studying volatility dynamics in the Israeli corporate bond market, where investor behavior and platform design jointly shape financial stability.

This study aims to fill the existing gap. We specifically focus on assessing various forecasting models for their efficacy in capturing the unique volatility features of the Israeli corporate bond market. Using the TASE bond indices, Tel-Bond 20 and Tel-Bond 60, we test

and analyze the forecasting performance of GARCH, EGARCH, GJR, and APARCH models, along with normal distribution density and asymmetric Student's *t*-distribution density. The outcomes of this research have broad implications, ranging from risk management and asset pricing to regulatory policy decisions.

Among the various GARCH models tested, we find that the GJR Student-*t* model is the most effective in capturing the unique characteristics of the Israeli corporate bond market. This model successfully captures asymmetric volatility clustering and high autocorrelation, providing more accurate volatility forecasts. The results also highlight the significant impact of investor sentiment on bond market volatility, with negative news leading to higher volatility than positive news. These findings suggest that behavioral frictions among retail investors are a major source of volatility in Israel's corporate bonds market, with price reactions more pronounced following negative sentiment, consistent with noise trader models (Foucault et al. 2011; Kumar and Lee 2006).

Our findings advance the literature on volatility forecasting in bond markets by highlighting that platforms with LOB mechanisms and high retail engagement display distinct volatility dynamics. Given the rising role of non-institutional investors in global bond markets and the trend toward transparent, electronic trading (Abudy and Shust 2023; Hadad 2025), these results have important implications for market design and financial stability. Specifically, they highlight that platform structure and investor composition can significantly amplify volatility through sentiment-driven trading. Accordingly, our study underscores the importance of applying well-established volatility models to underexplored market microstructures, where behavioral factors and market transparency jointly shape risk transmission and systemic resilience.

The study proceeds as follows: Section 2 provides a review of the theoretical impact of retail trading activity on the volatility of returns, and summarize the findings on corporate bond returns. Section 3 presents the data and the methodology of the models used in the study. Section 4 describes the estimation procedures and presents the forecasting results and comparisons. Section 5 concludes.

2. Literature Review

Theories of behavioral finance argue that asset prices may be affected by investors' psychological attributes. Shifts in sentiment, such as overconfidence, optimism, and wishful thinking, can significantly impact asset prices and their volatility (Barberis et al. 1998; Black 1986; Kyle 1985). Investor sentiment, broadly defined as investors' beliefs about future cash flows and investment risks (Baker and Wurgler 2006), may not be justified by fundamental news or facts. Moreover, it is costly and risky to bet against sentimental investors, meaning rational investors or arbitrageurs are not as aggressive in forcing prices back to fundamental values ls (Baker and Wurgler 2007). Thus, price anomalies may form when sentiment investors over (under) estimate return and underestimate (overestimate) risk, hence investing more on the risky (safer) asset cause a mispricing of the asset in relative to its risk-based fundamental (Baker and Wurgler 2006, 2007). Hence, noise traders, who often exhibit irrational behavior, potentially can contribute to increased market volatility, imposing higher risks on rational arbitrageurs and ultimately affecting asset prices (Foucault et al. 2011; Huerta-Sanchez and Escobari 2018).

Extensive research has documented empirical evidence of the impact of investor sentiment on stock return volatility. Lee et al. (2002) found that changes in sentiment are inversely correlated with the conditional volatility of U.S. stock market indexes. Yu and Yuan (2011) showed that high sentiment periods correlate with a positive tradeoff between the mean and variance of U.S. stock returns, suggesting greater influence of sentiment-driven investors. Verma and Verma (2007) demonstrated that sentiment-driven

retail investors significantly impact stock return volatility, with bullish sentiment having a greater effect than bearish sentiment. Other studies confirm that investor sentiment plays a significant role in international market volatility (Baker et al. 2012; Feldman and Liu 2017; Gong et al. 2022), highlighting that noise trading can contribute to increased market volatility.

Focusing on bond markets, fewer studies document the impact of shocks on U.S. Treasury bond volatilities, showing significant increases in bond market volatility on announcement days, which quickly subside as news is incorporated into prices (Christiansen 2000; Jones et al. 1998). Li and Engle (1998) demonstrate that news announcement shocks impact the volatility of U.S. Treasury bond futures, with volatility responding asymmetrically to these shocks. Other studies also document asymmetries in bond return volatilities (Cappiello et al. 2006; de Goeij and Marquering 2004, 2006). Additionally, Piazzesi (2005) shows that Federal Open Market Committee (FOMC) announcements are important determinants of bond market volatility, suggesting that bond investors underreact to information, implying that irrational behavior potentially impacts volatility in the debt market as well. Similarly, studies in other Asian markets have suggested the possibility of a leverage effect in bond yield volatility, implying for potential sentiment-driven impact on bond returns (BM et al. 2023; Mukherjee 2019; Rath 2023; Tan and Tian 2009). However, empirical evidence on volatility asymmetries in corporate bond markets remains limited, highlighting a clear gap in the literature that our study seeks to address.

If volatility is priced in the bond market, an anticipated increase in volatility would result in a higher required return in corporate bonds, which are generally perceived as more risky by investors (Acharya and Pedersen 2005; Dick-Nielsen et al. 2012; Friewald et al. 2012). Despite the well-documented impact of sentiment on stock and government bond market volatility, the influence of sentiment on corporate bond markets has received less attention. The few exceptions include studies that show U.S. bond yield spreads co-vary with sentiment, similar to stocks (Nayak 2010; Spyrou 2013). Specifically, information uncertainty and information asymmetry are found to be prices U.S corporate bonds yields spreads after controlling for credit ratings (Lu et al. 2010). Additionally, U.S. bonds appear underpriced (with high yields) during pessimistic periods and overpriced (with low yields) when optimism reigns (Nayak 2010). Similarly, Bethke et al. (2017) document that U.S. corporate bond investors exhibit a flight-to-quality when sentiment is low. In fact, the sentiment effect in U.S corporate bonds is found to spilled over from the stocks markets, through the activity of investors involved in capital structure arbitrage (Huang et al. 2015). This pattern is also evident in international corporate bond pricing and liquidity, which are generally affected by the same factors as the U.S. market (Goldstein and Namin 2023; Rath 2023). While these findings suggest that sentiment can indeed affect bond markets, they do not consider the impact on corporate bond volatility, which evidently changes over time in a pattern similar to stocks (Reilly et al. 2000). To the best of our knowledge, no prior study has examined sentiment-driven volatility in corporate bond markets within emerging economies—particularly using asymmetric GARCH models—thereby providing new insights into the behavioral dynamics of credit markets. Given the recent evidence of sentiment effects in bond pricing, more research is needed to fully understand this dynamic (Goldstein and Namin 2023).

Focusing on the Israeli market, several studies document a significant presence of retail trading activity, indicating that retail investors play a crucial role in enhancing market liquidity and efficiency (Abudy and Shust 2023; Abudy and Wohl 2018; Hadad 2025; Hadad and Kedar-Levy 2024). Hadad and Kedar-Levy (2024) show that the high presence of retail-sized investors has a positive impact on corporate bond returns and volatility. Using an EGARCH (1,1) model on the TA-35 (stock) Index and the Tel-Bond-20 (bond)

Index returns, they found that changes in market sentiment proxies, reflecting changes in risk expectations and investor sentiment, largely explain movements in the conditional volatility of both stock and bond market returns. This implies that investors in both stocks and corporate bonds should consider sentiment in their investment decisions, and that both asset classes may be attractive for speculative investors. However, their study focuses solely on the EGARCH (1,1) model and does not test several GARCH models to estimate market volatility efficiently.

Given the above-mentioned literature, we extend volatility models used in stock markets to corporate bonds. This pattern is particularly evident in the Israeli market, characterized by high retail trading activity (Abudy and Shust 2023; Hadad and Kedar-Levy 2024). Alberg et al. (2008) found that asymmetric GARCH models, such as EGARCH and GJR, were effective in capturing volatility dynamics in the Israeli stock market. Similarly, we aim to develop a volatility model that captures well-known stylized facts about conditional volatility, such as persistence, mean-reverting behavior, and asymmetric impacts of negative versus positive return innovations. We examine the estimation performance of various models, including GARCH, EGARCH, GJR, and APARCH, using different density functions: normal distribution and Student's *t*-distribution. Our focus is on the Tel-Bond 20 and Tel-Bond 60 indices, which reflect the Israeli corporate bond market. This methodology allows us to assess the suitability of different volatility forecasting models and emphasize the role of investor sentiment in explaining volatility patterns.

3. Data

The Israeli corporate bond market is one of the most liquid and transparent globally, supported by the LOB trading mechanism and dominated by active retail participation. According to the TASE 2024 Annual Report, the total market capitalization of corporate bonds reached approximately US$114.7 billion, including US$69.1 billion in CPI-linked bonds and US$45.6 billion in non-linked bonds.[1] The average daily trading volume was US$0.23 billion, marking a 6% increase from 2023. Notably, the Israeli public purchased US$4.7 billion in corporate bonds in 2024, offsetting net sales by foreign investors (US$1.9 billion) and long-term institutional investors (US$2.7 billion). These figures highlight the prominent role of retail investors and the distinctive exchange-based bond trading environment, making TASE an ideal setting for examining sentiment-driven volatility dynamics in the bond market.

Our data consist of 738 daily observations of the Tel-Bond 20 Index prices and 738 daily observations of the Tel-Bond 60 Index prices, covering the period from 1 July 2019, to 30 June 2022. These data were obtained from the TASE website[2]. The rationale for choosing this specific dataset lies in its relevance for modeling volatility during critical economic periods, notably the COVID-19 pandemic and the inflation period of 2022 in Israel. During the COVID-19 pandemic, the corporate bonds market in Israel experienced significant volatility due to economic uncertainty and market disruptions (Hadad and Kedar-Levy 2024). Additionally, during inflationary times, bonds generally exhibit higher volatility (Campbell et al. 2020; Kang and Pflueger 2015), making this period particularly valuable for studying volatility patterns.

The Tel-Bond 20 and Tel-Bond 60 indices are key benchmarks in the Israeli corporate bond market, representing the top 20 and 60 fixed-coupon and CPI-linked corporate bonds, respectively, based on market capitalization. According to the TASE website, the threshold conditions for inclusion in these indices require a minimum rating from Israeli rating companies "Midroog" and "Maalot" of A or A3. Bonds meeting these criteria are included in the index, while those falling below the exit rating are removed. Typically, the rating is based on the average market value. The entry and exit ratings for the Tel-Bond 20

index are 16 and 24, respectively. As for the Tel-Bond 60 index, no specific entry and exit ratings have been determined due to its inclusion of a broader range of companies. These indices offer a comprehensive view of market behavior, making them excellent choices for studying volatility.

Table 1 shows descriptive statistics for the daily log returns of the Tel-Bond 20 and Tel-Bond 60 indices, calculated as the natural logarithm of the ratio of consecutive prices. The mean and median returns for both indices are close to zero, indicating that the average daily returns are quite small. The standard deviation is slightly higher for the Tel-Bond 20 index (0.0039) compared to the Tel-Bond 60 index (0.0036), suggesting that the Tel-Bond 20 is slightly more volatile. This result may be associated with the concentration of higher market capitalization bonds in the Tel-Bond 20 index, leading to greater price movements due to higher trading volumes and more significant investor reactions to market news and events. In particular, retail investors, who are more prone to sentiment-driven trading (Baker and Wurgler 2007; Edwards et al. 2007), may have a larger impact on the Tel-Bond 20 index, leading to increased volatility compared to the broader and more diversified Tel-Bond 60 index.

Table 1. Descriptive statistics.

Index	Mean	Median	Max	Min	Std. Dev.	Skewness	Kurtosis
Tel-Bond-20	0.000012	0.0002	0.0340	−0.0258	0.0039	0.6443	25.6379
Tel-Bond-60	0.000015	0.0002	0.0312	−0.0234	0.0036	0.5124	26.3665

Notes: This table presents descriptive statistics of the daily log-returns for the Tel-Bond-20 and Tel-Bond-60 indices traded on the Tel Aviv Stock Exchange (TASE) during the sample period.

The maximum and minimum values show that the Tel-Bond 20 index has had both higher peaks and deeper troughs than the Tel-Bond 60 index. Both indices show excess kurtosis and positive skewness, indicating that there are more extreme positive returns than extreme negative returns. The higher skewness in the Tel-Bond 20 index (0.6443) compared to the Tel-Bond 60 index (0.5124) suggests that the Tel-Bond 20 index has experienced more frequent large positive returns, which may be associated with sentiment-driven trading activity. The high kurtosis values for both indices indicate that the returns distribution has heavy tails and sharp peaks compared to a normal distribution, suggesting the presence of outliers in the time series.

Table 2 reports the results of the Jarque–Bera test for normality (Jarque and Bera 1980) and the ARCH LM test for stationarity. The Jarque–Bera test results are highly significant for both index returns, leading to the rejection of the null hypothesis of normality. The ARCH LM test results are also highly significant for both indices, indicating that the returns of both indices exhibit volatility clustering, implying that periods of high volatility tend to cluster together. This pattern can also be attributed to retail-trading activity, which is known to induce such clustering in financial markets. Unreported Augmented Dickey–Fuller (ADF) (Dickey and Fuller 1981) and Phillips Perron (Phillips and Perron 1988) stationarity tests results also show high significance, suggesting stationarity in the time series. Overall, these findings support the necessity of employing GARCH modeling to capture the volatility dynamics accurately in our corporate bond indices.

Table 2. Jarque–Bera and LM test.

Index	Jarque–Bera	LM Statistic
Tel-Bond-20	15,788.27 ***	59.1581 ***
Tel-Bond-60	16,798.75 ***	91.8002 ***

Notes: *** indicates significance at the 1% level. Jarque–Bera tests for normality; LM tests for ARCH effects. Both reject the null hypotheses, supporting the use of GARCH models.

4. Methodology

In this section we succinctly describe the GARCH models used in the study. We use the daily returns of both indices and model the mean equation as follows:

$$r_t = \mu + \varepsilon_t, \tag{1}$$

where r_t represents the return of our corporate bond index at time t, μ is a constant term, and ε_t the error term used to model the conditional volatility in the various GARCH models.

First, we implemented the GARCH model, which imposes nonlinear restrictions (Park 2002). This model improves upon the Auto-Regressive Conditional Heteroskedasticity (ARCH) models by Engle (1982) by adding a more flexible lag structure. The GARCH (1,1) variance estimation process is given by:

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

where α, β and $\omega = \alpha_0$ are the model parameters.

While the GARCH model is effective, it has limitations such as not accounting for the sign (positive or negative) of delayed innovations and exhibiting excess kurtosis of the residuals (Park 2002). To overcome these limitations, Bollerslev (1987) proposed using Student's t-distribution, which can capture conditional leptokurtosis separately from conditional heteroskedasticity. Several empirical studies have employed the generalized t-distribution to capture the skewness and leverage effects of daily returns and to address kurtosis and skewness limitations (Hansen 1994; Harris et al. 2004). However, the GARCH model does not handle the asymmetric effect, potentially biasing its estimation of conditional volatility (Villar-Rubio et al. 2023).

To handle the asymmetric effect, we explore nonlinear asymmetric models, including the Exponential GARCH (EGARCH) (Nelson 1991), the GJR model by (Glosten et al. 1993) and the Asymmetric Power ARCH (APARCH) model (Ding et al. 1993). These models consider the magnitudes and signs of shocks to conditional variance and explain the leverage effect (Lama et al. 2015). Given the asymmetric response of investors to good and bad news (Barberis et al. 1998; Hadad and Kedar-Levy 2024; Verma and Verma 2007), these models may be more suitable to model the volatility.

We apply the EGARCH model (Nelson 1991), which identifies asymmetric effects on conditional volatility. This model ensures that conditional variance remains positive by specifying it in logarithmic form, thus avoiding restrictions on the model's coefficients (López-Cabarcos et al. 2021; Piñeiro-Chousa et al. 2022). The logarithmic conditional variance of the corporate bond index is modeled using the EGARCH (1,1) model:

$$\log\left(\sigma_t^2\right) = \omega + \alpha\left[\frac{|\varepsilon_{t-1}|}{\sigma_{t-1}} - \frac{\sqrt{2}}{\pi}\right] + \gamma\left(\frac{\varepsilon_{t-1}}{\sigma_{t-1}}\right) + \beta log\left(\sigma_{t-1}^2\right), \tag{3}$$

where σ_t^2 is the conditional variance of the error term of the mean equation of the index, ε_{t-1} is the first-order lag from (1), and α represents the symmetric effect of the general

autoregressive model. ω is a constant and γ coefficient captures the asymmetric effect of innovations on the volatility of the index returns, when $\gamma < 0$ negative news generates higher volatility than positive news. β measures the stationary of the conditional volatility.

However, the EGARCH model has several limitations, such as capturing the leverage effect depending on the signs of the parameters and being highly sensitive to initial values. Additionally, the log-transformation in EGARCH models ensures positivity of the conditional variance but assumes a multiplicative effect, which might not always align with the actual data characteristics (Alberg et al. 2008). To address this, Glosten et al. (1993) extended the standard GARCH model by incorporating asymmetric effects of positive and negative shocks on volatility. For one lag in the return of the corporate bond index and variance, its conditional variance is modeled using the GJR (1,1) model as follows:

$$\sigma_t^2 = \omega + \alpha\varepsilon_{t-1}^2 + \gamma I_{t-1}\varepsilon_{t-1}^2 + \beta\sigma_{t-1}^2, \tag{4}$$

where I_{t-1} is an indicator function that is equal to 1 if $\varepsilon_{t-1} < 0$ and zero otherwise, and $\omega, \alpha, \beta, \gamma$ are the model parameters, while $\alpha, \beta \geq 0$, and $\omega > 0$. The term $\gamma I_{t-1}\varepsilon_{t-1}^2$ adds an extra component to the variance equation when the lagged error term is negative, capturing the phenomenon where negative shocks have a larger impact on volatility than positive shocks of the same magnitude (Glosten et al. 1993). While the GJR model effectively captures asymmetry through the indicator function, it does not allow for varying the power of the shocks. Furthermore, the GJR-GARCH model does not explicitly account for heavy-tailed distributions of returns, which are evident in time series of returns (Denes et al. 2023; Opschoor et al. 2018).

To address this issues, we also utilize the APARCH(1,1) model of Ding et al. (1993), which provides a more flexible approach by allowing for different powers of the absolute returns and asymmetry, by introducing a power parameter (δ) that can be optimized to better fit the data. Our APARCH(1,1) model is given by

$$\sigma_t^\delta = \omega + \alpha_1(|\varepsilon_{t-1}| - \gamma_1\varepsilon_{t-1})^\delta + \beta_1\sigma_{t-1}^\delta, \tag{5}$$

where $-1 < \gamma < 1$, $\delta > 0$ and ω, α, γ, β and δ are parameters to be estimated. The APARCH model provides a more flexible framework for modeling asymmetric effects and can better handle heavy-tailed distributions and varying volatility patterns, making it a more comprehensive tool for analyzing financial time series. Furthermore, Alberg et al. (2008) show that fat-tail distributions are better suited for modeling returns in the Israeli market, underscoring the necessity of the APARCH (1,1) model.

Since our returns time series deviate from the normality assumption, a Maximum Likelihood (ML) method is employed to estimate the models' parameters. We use the quasi-maximum likelihood estimator (QMLE) to obtain maximum likelihood estimates of the various GARCH model parameters. This method maximizes the Gaussian log-likelihood function of the multivariate normal distribution and results in consistent and asymptotic normality of the estimated parameters (Allen and McAleer 2018; Asai et al. 2021). Additionally, we utilize the Broyden–Fletcher–Goldfarb–Shanno (BFGS) optimization algorithm for estimation to select the best-fitted model. This algorithm minimizes the likelihood function (Mahmood and Khan 2020). Following Alberg et al. (2008), we apply the QMLE method under the normal and Student's *t*-distributions to select the best-fitted model.

Lastly, we select the optimal model by measuring several standard criteria to determine the most adequate specification. The best model was selected using the Akaike Information Criterion (AIC), Schwarz Information Criterion (SIC), and Hannan–Quinn Criterion (HQIC) statistics. These information criteria indicate how well the estimated model fits the data compared to other models (Shittu and Asemota 2009), allowing us to

better capture volatility clustering behavior for efficient risk management strategies and better-informed portfolio allocation decisions.

5. Results

Table 3 summarizes the results of the GARCH (1,1) model on the Tel-Bonds indices daily returns.

Table 3. GARCH (1,1) model results.

Estimation Results of the Variance Equation:	$\sigma_t^2=\omega+\alpha\varepsilon_{t-1}^2+\beta\sigma_{t-1}^2$			
Variable	**Normal**		**Student's *t***	
	Tel-Bond-20	**Tel-Bond-60**	**Tel-Bond-20**	**Tel-Bond-60**
ω	2.43×10^{-7} ***	2.04×10^{-7} ***	2.50×10^{-7} ***	2.23×10^{-7} ***
α	0.201152 ***	0.199625 ***	0.223436 ***	0.231301 ***
β	0.779118 ***	0.781201 ***	0.765361 ***	0.758634 ***
Adjusted R-squared	0.040581	0.040827	0.039660	0.039538
Log likelihood	3411.521	3478.030	3424.784	3495.102
Akaike Info Criterion	−9.256851	−9.437582	−9.290173	−9.481255
Schwarz Criterion	−9.225593	−9.406324	−9.252663	−9.443745
Hannan–Quinn Criterion	−9.244796	−9.425527	−9.275707	−9.466788

Notes: The table reports GARCH(1,1) estimates under both normal and Student's *t* error distributions. *** indicate significance at the 1% level.

The GARCH coefficients are positive and highly significant, indicating that current volatility is highly sensitive to past information, both errors (α) and volatility (β). The α coefficient, representing model errors, is approximately 0.2, while β coefficient, representing past conditional variance, is approximately 0.8. These results suggest that current volatility of bond returns is more influenced by past volatility than by past shocks. Furthermore, the information criteria (AIC, SBIC, and HQIC) have minimal values for Student's *t*-distribution, indicating that it is a better fit model. These results align with BM et al. (2023), who noted the superiority of Student's *t*-distribution for modeling financial time series with heavy tails.

Table 4 show estimation results for the EGARCH (1,1) model. The EGARCH model results, particularly with Student's *t*-distribution, indicate that all coefficients are highly significant. The preference for Student's *t*-distribution, supported by lower AIC, SIC, and HQIC values, aligns with Yong et al. (2021), who identified this distribution as optimal for EGARCH (1,1) modeling of other Asian markets during periods of market turbulence. For this distribution, we document a highly significant α coefficient, indicating the symmetric effect of past errors on volatility, with values as approximately of 0.3. The β coefficient, which is around 0.9, indicates long-term persistence in bond market volatility.

The asymmetry term shows a negative γ coefficient around -0.1 for both Tel-Bond 20 and Tel-Bond 60 returns, implying that negative shocks have a more significant impact on volatility than positive shocks of the same magnitude, demonstrating a leverage effect. This is consistent with findings by Hadad and Kedar-Levy (2024) in the context of the Israeli corporate bonds market during the COVID-19 pandemic. These results underscore the importance of accounting for asymmetry in volatility modeling, as negative news tends to amplify volatility more than positive news, reflecting the behavioral biases of investors.

Table 5 summarizes the results of the GJR (1,1) model on the indices daily returns. The GJR model results, particularly with Student's *t*-distribution, indicate that the coefficients are positive and highly significant. The β coefficient of approximately 0.8 indicates the persistence of conditional volatility. The positive γ coefficient confirms the presence

of a leverage effect in the Israeli bond market during the COVID-19 period, indicating that negative shocks have a larger impact on volatility than positive shocks of the same magnitude. This is especially evident when using Student's *t*-distribution, where the α coefficient remains highly significant, unlike in the normal distribution where it is insignificant.

Table 4. EGARCH (1,1) model results.

Estimation Results of the Variance Equation: $log(\sigma^2)=\omega+\alpha\left[\frac{\lvert\varepsilon_{t-1}\rvert}{\sigma_{t-1}}-\frac{\sqrt{2}}{\pi}\right]+\gamma\left(\frac{\varepsilon_{t-1}}{\sigma_{t-1}}\right)+\beta log\left(\sigma^2_{t-1}\right)$				
Variable	**Normal**		**Student's *t***	
	Tel-Bond-20	**Tel-Bond-60**	**Tel-Bond-20**	**Tel-Bond-60**
ω	−11.80657 ***	−11.58546 ***	−0.680550 ***	−0.721949 ***
α	1.107948 ***	1.153876 ***	0.321036 ***	0.334494 ***
β	0.055067	0.093176	0.963587 ***	0.961363 ***
γ	−0.019139	−0.103186	−0.101851 ***	−0.092911 ***
Adjusted R-squared	0.039736	0.042865	0.042187	0.041403
Log likelihood	3260.014	3336.649	3425.896	3495.751
Akaike Info Criterion	−8.842429	−9.050678	−9.290478	−9.480302
Schwarz Criterion	−8.804919	−9.013168	−9.246717	−9.436540
Hannan–Quinn Criterion	−8.827963	−9.036212	−9.273601	−9.463424

Note: The table reports EGARCH(1,1) estimates for Tel-Bond-20 and Tel-Bond-60 indices returns under both normal and Student's *t* error distributions. γ captures asymmetry (leverage effects). *** indicate significance at the 1% level.

Table 5. GJR (1,1) model results.

Estimation Results of the Variance Equation: $\sigma^2_t=\omega+\alpha\varepsilon^2_{t-1}+\gamma I_{t-1}\varepsilon^2_{t-1}+\beta\sigma^2_{t-1}$				
Variable	**Normal**		**Student's *t***	
	Tel-Bond-20	**Tel-Bond-60**	**Tel-Bond-20**	**Tel-Bond-60**
ω	2.60×10^{-7} ***	2.23×10^{-7} ***	2.62×10^{-7} ***	2.32×10^{-7} ***
α	0.072221	0.080696 *	0.087298 **	0.104001 **
γ	0.194331 ***	0.184904 ***	0.203293 ***	0.189128 ***
β	0.798634 ***	0.794868 ***	0.784243 ***	0.775620 ***
Adjusted R-squared	0.042801	0.042593	0.041522	0.040910
Log likelihood	3419.856	3485.336	3430.154	3499.254
Akaike Info Criterion	−9.276783	−9.454718	−9.302048	−9.489822
Schwarz Criterion	−9.239273	−9.417208	−9.258286	−9.446060
Hannan–Quinn Criterion	−9.262317	−9.440252	−9.285170	−9.472945

Notes: This table reports GJR-GARCH(1,1) estimates for Tel-Bond-20 and Tel-Bond-60 indices return under both Normal and Student's *t* distributions. γ captures asymmetric effects of negative shocks (leverage effects).***, ** and * indicate significance at the 1%, 5% and 10% level, respectively.

The superior performance of Student's *t*-distribution for the GJR model is further underscored by the minimal AIC, SIC, and HQIC values, highlighting its better fit for the data. This finding is consistent with Alberg et al. (2008), who found that models assuming a Student's *t*-distribution provide a better fit for financial time series data in the Israeli stock market. This underscores the importance of using a distribution that can capture the heavy tails and excess kurtosis often observed in financial returns, making Student's *t*-distribution a robust choice for modeling bond market volatility during periods of market turbulence.

Finally, we estimated the conditional variance of the bond yield indices using the Asymmetric Power ARCH (APARCH) model. This model, which has the ability to generate many ARCH models by varying the parameters, couples the flexibility of a varying exponent with the asymmetry coefficient. Table 6 summarizes the results of the APARCH (1,1) model.

Table 6. APARCH (1,1) model results.

Estimation Results of the Variance Equation: $\sigma_t^\delta = \omega + \sum_{i=1}^{q} \alpha_i (\lvert \varepsilon_{t-i} \rvert - \gamma_i \varepsilon_{t-i})^\delta + \sum_{j=1}^{p} \beta_j \sigma_{t-j}^\delta$		
Variable	**Student's *t***	
	Tel-Bond-20	**Tel-Bond-60**
ω	2.29×10^{-8}	4.29×10^{-8}
α	0.166892 ***	0.181514 ***
γ	0.263686 ***	0.236297 **
β	0.764342 ***	0.762282 ***
δ	2.398791 ***	2.273031 ***
Adjusted R-squared	0.041453	0.04960
Log likelihood	3430.362	3499.351
Akaike Info Criterion	−9.299896	−9.487368
Schwarz Criterion	−9.249883	−9.437354
Hannan–Quinn Criterion	−9.280608	−9.468080

Note: This table presents the estimated parameters of the Asymmetric Power ARCH (APARCH) model for the Tel-Bond-20 and Tel-Bond-60 indices return using Student's *t* distribution. γ and δ capture the leverage effect and the power term, respectively, enhancing model flexibility *** and ** indicate significance at the 1% and 5% level, respectively.

For both indices, convergence could not be reached with the APARCH model and a normal distribution. Therefore, we used Student's *t* version of the model. The results show that the coefficients α and β are positive and statistically significant. The power coefficient δ is positive and significant as well, and not equal to 2, establishing that it is not a standard GARCH model (Ding et al. 1993). The asymmetry coefficient γ is positive and significant, indicating the existence of a leverage effect where negative news increases volatility more than positive news of the same magnitude. The significant α coefficient implies that past errors significantly affect current volatility, while the high β coefficient indicates long-term volatility persistence.

Overall, the results show the presence of a leverage effect, where negative shocks have a more significant impact on volatility than positive shocks. This effect is consistently observed across the EGARCH, GJR and APARCH models, indicating that investor sentiment and reactions to negative news play a crucial role in driving bond market volatility. These findings highlight the importance of considering asymmetry and heavy-tailed distributions in modeling financial time series, especially in markets with significant retail trading activity.

The economic implications of these findings are significant, suggesting that investors should be cautious during periods of market stress and that portfolio allocation strategies should account for potential increases in volatility. We attribute these findings to the evolution of the unique trading platform, which improved market practices and trading behavior. This insight is crucial for developing more robust risk management strategies and ensuring a resilient financial system

Lastly, to test the accuracy of volatility forecasting among the different models with distribution assumptions, we compare several standard criteria: AIC, SBIC, HQIC and the Log-Like value. Given that the results across all models indicate the superior fit of Student's *t*-distribution over the normal distribution, we focus on comparing Student's *t*-distribution

models to better capture the volatility dynamics of the Tel-Bond 20 and Tel-Bond 60 indices. Results for Tel-Bond 20 and Tel-Bond 60 are presented in Table 7.

Table 7. Comparison between the models for the Tel Bond 20.

	Student's *t* GARCH	**Student's *t* EGARCH**	**Student's *t* GJR**	**Student's *t* APARCH**
		Tel-Bond-20		
AIC	−9.290	−9.290	−9.302	−9.299
Schwarz	−9.253	−9.247	−9.258	−9.225
Hannan–Quinn	−9.276	−9.274	−9.285	−9.280
		Tel-Bond-60		
AIC	−9.481	−9.480	−9.490	−9.487
Schwarz	−9.444	−9.437	−9.446	−9.437
Hannan–Quinn	−9.467	−9.463	−9.473	−9.468

Notes: This table compares the performance of the four volatility models for the Tel-Bond-20 index and Tel-Bond 60 index returns. All models are estimated with Student's *t*-distribution. Lower AIC, Schwarz, and Hannan–Quinn information criteria indicate better the model fit.

The comparison of the different models for the Tel-Bond 20 and Tel-Bond 60 indices, as summarized in Table 7, demonstrates that the GJR model with a Student's *t*-distribution provides the best fit for the data. For the Tel-Bond 20 index, the GJR model outperformed other models with an AIC of −9.302, an SBIC of −9.258, and an HQIC of −9.285. Similarly, for the Tel-Bond 60 index, the GJR model achieved the lowest AIC of −9.490, an SBIC of −9.446, and an HQIC of −9.473. These results indicate that the GJR model with a Student's *t*-distribution better captures the volatility dynamics of both indices compared to the GARCH, EGARCH, and APARCH models.

While these findings are consistent with previous studies documenting the outperformance of the GJR model in estimating stock returns volatility (Liu and Hung 2010), they contrast with findings for the Israeli market. Specifically, Alberg et al. (2008) found the EGARCH model to be the most successful for measuring conditional variance and forecasting the Israeli stock indices. Furthermore, Hadad and Kedar-Levy (2024) documented that the EGARCH(1,1) model had the best fit for the Tel-Bond 20 index returns. However, it should be noted that these studies primarily focused on outdated data. For instance, the sample in Hadad and Kedar-Levy (2024) ranges from 2000 to 2019, thus not considering the significant impact of COVID-19 and the rise in inflation, which evidently impact the Israeli corporate bonds market. Given that economic uncertainty and inflationary times exhibit higher volatility in corporate bonds (Bethke et al. 2017; Campbell et al. 2020; Engelberg et al. 2018; Kang and Pflueger 2015), this should favor the GJR results.

6. Conclusions

This study examines conditional volatility in the Israeli corporate bond market using GARCH family models. Our findings reveal that the GJR-GARCH model with a Student's *t*-distribution most effectively captures the asymmetric volatility behavior of bond returns, highlighting the presence of a leverage effect and sentiment-driven volatility. These findings reflect that negative shocks lead to disproportionately higher volatility, highlighting increased sensitivity to downside risk. These dynamics are especially relevant in transparent and retail-dominated markets like Israel, where information is rapidly incorporated into prices. Such volatility asymmetry may distort risk premia, widen spreads, and impair market functioning during stress episodes, and thus accurate volatility modeling becomes essential for risk management, asset pricing, and regulatory oversight.

These findings are particularly relevant for the Israeli market, characterized by high levels of transparency and significant retail participation. Prior work shows that market transparency enhances liquidity and narrows bid-ask spreads (Bessembinder and Maxwell 2008; Cici et al. 2011; Goldstein et al. 2007), but also increases market responsiveness to new information. In such settings, retail-driven sentiment plays a larger role in shaping price volatility. Indeed, Baker and Wurgler (2006, 2007) highlight how retail sentiment intensifies volatility and exacerbates deviations from fundamentals—especially in less institutionalized markets. This supports recent findings by Goldstein and Namin (2023) and Wang (2023) that bond markets with active retail investors exhibit more pronounced volatility in response to sentiment shifts.

Our results reinforce the suitability of ARCH-type models for bond volatility analysis, in line with Reilly et al. (2000), and underscore the importance of incorporating asymmetry in volatility forecasts for transparent, sentiment-sensitive markets. Using appropriate models such as GJR-GARCH can improve risk management, enhance price discovery, and support financial stability initiatives in such environments.

Nonetheless, the study is limited to a univariate GARCH framework, subjective model selection, and a sample period that reflects a turbulent economic time (the COVID-19 pandemic), which may not capture the full diversity of market conditions and could limit the generalizability of the findings to more stable periods. Moreover, the localized nature of the Israeli corporate bond market, despite its unique structural strengths, further constrains the broader applicability of our results. Future research should extend the analysis across different corporate bond sectors within Israel, consider multivariate and regime-switching approaches, and evaluate how platform design and investor composition affect volatility patterns. Crucially, we recommend applying this framework to other emerging corporate bond markets with similar structural traits, namely high transparency and strong retail investor presence, to enable comparative insights into sentiment-driven volatility dynamics. Such extensions can guide both local and international policymakers in designing interventions that mitigate systemic risk under uncertainty.

Author Contributions: Conceptualization, E.H. and R.Y.; methodology, E.H.; software, A.M.F.; validation, A.M.F.; formal analysis, A.M.F.; investigation, A.M.F.; resources, R.Y.; data curation, A.M.F.; writing—original draft preparation, A.M.F.; writing—review and editing, E.H.; visualization, A.M.F.; supervision, E.H. and R.Y.; project administration, E.H. and R.Y.; funding acquisition, R.Y. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The data presented in this study are openly available in https://www.tase.co.il/en (accessed on 4 July 2022).

Conflicts of Interest: The authors declare no conflicts of interest.

Notes

1 https://content.tase.co.il/media/xt0h0tff/1740653403_tase_2024annualreviewtenglsih.pdf?guid=00e35b48-abf9-4d19-b27b-cedfcea0a796 (accessed on 10 April 2025).

2 https://www.tase.co.il/en (accessed on 10 April 2025).

References

Abudy, Menachem Meni, and Avi Wohl. 2018. Corporate bond trading on a limit order book exchange. *Review of Finance* 22: 1413–40. [CrossRef]

Abudy, Menachem Meni, and Efrat Shust. 2023. Does market design contribute to market stability? Indications from a corporate bond exchange during the COVID-19 crisis. *Journal of Economics and Business* 123: 106105. [CrossRef]

Acharya, Viral V., and Lasse Heje Pedersen. 2005. Asset pricing with liquidity risk. *Journal of Financial Economics* 77: 375–410. [CrossRef]

Alberg, Dima, Haim Shalit, and Rami Yosef. 2008. Estimating stock market volatility using asymmetric GARCH models. *Applied Financial Economics* 18: 1201–8. [CrossRef]

Allen, David E., and Michael McAleer. 2018. Theoretical and empirical differences between diagonal and full BEKK for risk management. *Energies* 11: 1627. [CrossRef]

Asai, Manabu, Chia-Lin Chang, Michael McAleer, and Laurent Pauwels. 2021. Asymptotic and finite sample properties for multivariate rotated garch models. *Econometrics* 9: 21. [CrossRef]

Attarzadeh, Amirreza, and Mehmet Balcilar. 2022. On the Dynamic Connectedness of the Stock, Oil, Clean Energy, and Technology Markets. *Energies* 15: 1893. [CrossRef]

Bai, Jennie, Turan G. Bali, and Quan Wen. 2021. Is there a risk-return tradeoff in the corporate bond market? Time-series and cross-sectional evidence. *Journal of Financial Economics* 142: 1017–37. [CrossRef]

Baker, Malcolm, and Jeffrey Wurgler. 2006. Investor sentiment and the cross-section of stock returns. *The Journal of Finance* 61: 1645–80. [CrossRef]

Baker, Malcolm, and Jeffrey Wurgler. 2007. Investor sentiment in the stock market. *Journal of Economic Perspectives* 21: 129–51. [CrossRef]

Baker, Malcolm, Jeffrey Wurgler, and Yu Yuan. 2012. Global, local, and contagious investor sentiment. *Journal of Financial Economics* 104: 272–87. [CrossRef]

Barberis, Nicholas, Andrei Shleifer, and Robert Vishny. 1998. A model of investor sentiment. *Journal of Financial Economics* 49: 307–43. [CrossRef]

Bessembinder, Hendrik, and William Maxwell. 2008. Markets: Transparency and the corporate bond market. *Journal of Economic Perspectives* 22: 217–34. [CrossRef]

Bethke, Sebastian, Monika Gehde-Trapp, and Alexander Kempf. 2017. Investor sentiment, flight-to-quality, and corporate bond comovement. *Journal of Banking & Finance* 82: 112–32. [CrossRef]

Black, Fischer. 1986. Noise. *The Journal of Finance* 41: 528–43. [CrossRef]

BM, Lithin, Suman Chakraborty, Vishwanathan Iyer, Nikhil MN, and Sanket Ledwani. 2023. Modelling asymmetric sovereign bond yield volatility with univariate GARCH models: Evidence from India. *Cogent Economics & Finance* 11: 2189589. [CrossRef]

Bollerslev, Tim. 1987. A Conditionally Heteroskedastic Time Series Model for Speculative Prices and Rates of Return. *The Review of Economics and Statistics* 69: 542–47. [CrossRef]

Brown, Gregory W., and Michael T. Cliff. 2004. Investor sentiment and the near-term stock market. *Journal of Empirical Finance* 11: 1–27. [CrossRef]

Campbell, John Y., Carolin Pflueger, and Luis M. Viceira. 2020. Macroeconomic drivers of bond and equity risks. *Journal of Political Economy* 128: 3148–85. [CrossRef]

Cappiello, Lorenzo, Robert F. Engle, and Kevin Sheppard. 2006. Asymmetric dynamics in the correlations of global equity and bond returns. *Journal of Financial Econometrics* 4: 537–72. [CrossRef]

Choi, Jaewon, and Yongjun Kim. 2018. Anomalies and market (dis)integration. *Journal of Monetary Economics* 100: 16–34. [CrossRef]

Christiansen, Charlotte. 2000. Macroeconomic announcement effects on the covariance structure of government bond returns. *Journal of Empirical Finance* 7: 479–507. [CrossRef]

Cici, Gjergji, Scott Gibson, and John J. Merrick, Jr. 2011. Missing the marks? Dispersion in corporate bond valuations across mutual funds. *Journal of Financial Economics* 101: 206–26. [CrossRef]

Clayton, Jim, David C. Ling, and Andy Naranjo. 2009. Commercial real estate valuation: Fundamentals versus investor sentiment. *The Journal of Real Estate Finance and Economics* 38: 5–37. [CrossRef]

de Goeij, Peter, and Wessel Marquering. 2004. Modeling the Conditional Covariance Between Stock and Bond Returns: A Multivariate GARCH Approach. *Journal of Financial Econometrics* 2: 531–64. [CrossRef]

de Goeij, Peter, and Wessel Marquering. 2006. Macroeconomic announcements and asymmetric volatility in bond returns. *Journal of Banking & Finance* 30: 2659–80. [CrossRef]

Denes, Matthew, Sabrina T. Howell, Filippo Mezzanotti, Xinxin Wang, and Ting Xu. 2023. Investor Tax Credits and Entrepreneurship: Evidence from U.S. States. *The Journal of Finance* 78: 2621–71. [CrossRef]

Dickey, David A., and Wayne A. Fuller. 1981. Likelihood Ratio Statistics for Autoregressive Time Series with a Unit Root. *Econometrica: Journal of the Econometric Society* 49: 1057–72. [CrossRef]

Dick-Nielsen, Jens, Peter Feldhütter, and David Lando. 2012. Corporate bond liquidity before and after the onset of the subprime crisis. *Journal of Financial Economics* 103: 471–92. [CrossRef]

Ding, Zhuanxin, Clive W. J. Granger, and Robert F. Engle. 1993. A long memory property of stock market returns and a new model. *Journal of Empirical Finance* 1: 83–106. [CrossRef]

Ederington, Louis, Wei Guan, and Lisa Zongfei Yang. 2015. Bond market event study methods. *Journal of Banking & Finance* 58: 281–93. [CrossRef]

Edwards, Amy K., Lawrence E. Harris, and Michael S. Piwowar. 2007. Corporate bond market transaction costs and transparency. *The Journal of Finance* 62: 1421–51. [CrossRef]

Engelberg, Joseph, R. David McLean, and Jeffrey Pontiff. 2018. Anomalies and News. *The Journal of Finance* 73: 1971–2001. [CrossRef]
Engle, Robert F. 1982. Autoregressive Conditional Heteroscedasticity with Estimates of the Variance of United Kingdom Inflation. *Econometrica: Journal of the Econometric Society* 50: 987–1007. [CrossRef]
Feldman, Todd, and Shuming Liu. 2017. Contagious investor sentiment and international markets. *Journal of Portfolio Management* 43: 125. [CrossRef]
Foucault, Thierry, David Sraer, and David J. Thesmar. 2011. Individual Investors and Volatility. *The Journal of Finance* 66: 1369–406. [CrossRef]
Friewald, Nils, Rainer Jankowitsch, and Marti G. Subrahmanyam. 2012. Illiquidity or credit deterioration: A study of liquidity in the US corporate bond market during financial crises. *Journal of Financial Economics* 105: 18–36. [CrossRef]
Glosten, Lawrence R., Ravi Jagannathan, and David E. Runkle. 1993. On the Relation between the Expected Value and the Volatility of the Nominal Excess Return on Stocks. *The Journal of Finance* 48: 1779–801. [CrossRef]
Goldstein, Michael A., and Elmira Shekari Namin. 2023. Corporate bond liquidity and yield spreads: A review. *Research in International Business and Finance* 65: 101925. [CrossRef]
Goldstein, Michael A., Edith S. Hotchkiss, and Erik R. Sirri. 2007. Transparency and liquidity: A controlled experiment on corporate bonds. *The Review of Financial Studies* 20: 235–73. [CrossRef]
Gong, Xiao-Li, Jian-Min Liu, Xiong Xiong, and Wei Zhang. 2022. Research on stock volatility risk and investor sentiment contagion from the perspective of multi-layer dynamic network. *International Review of Financial Analysis* 84: 102359. [CrossRef]
Graham, John R., Mark T. Leary, and Michael R. Roberts. 2015. A century of capital structure: The leveraging of corporate America. *Journal of Financial Economics* 118: 658–83. [CrossRef]
Gur-Gershgoren, Gitit, Haim Kedar-Levy, and Elroi Hadad. 2020. Deep-Market by IAS-19: A Unified Cross-Country Approach for Discount Rate Selection. *Multinational Finance Journal* 24: 119–54.
Hadad, Elroi. 2025. Does trading mechanism shape cross-market integration? Evidence from stocks and corporate bonds on the Tel Aviv Stock Exchange. *Journal of Economics, Finance and Administrative Science* 30: 169–88. [CrossRef]
Hadad, Elroi, and Haim Kedar-Levy. 2024. The impact of retail investor sentiment on the conditional volatility of stocks and bonds: Evidence from the Tel-Aviv stock exchange. *International Review of Economics & Finance* 89: 1303–13. [CrossRef]
Hansen, Bruce E. 1994. Autoregressive Conditional Density Estimation. *International Economic Review* 35: 705–30. [CrossRef]
Harris, Richard D. F., C. Coskun Küçüközmen, and Fatih Yilmaz. 2004. Skewness in the conditional distribution of daily equity returns. *Applied Financial Economics* 14: 195–202. [CrossRef]
Huang, Jing-Zhi, Marco Rossi, and Yuan Wang. 2015. Sentiment and corporate bond valuations before and after the onset of the credit crisis. *The Journal of Fixed Income* 25: 34. [CrossRef]
Huerta-Sanchez, Daniel, and Diego Escobari. 2018. Changes in sentiment on REIT industry excess returns and volatility. *Financial Markets and Portfolio Management* 32: 239–74. [CrossRef]
Jarque, Carlos M., and Anil K. Bera. 1980. Efficient tests for normality, homoscedasticity and serial independence of regression residuals. *Economics Letters* 6: 255–59. [CrossRef]
Jones, Charles M., Owen Lamont, and Robin L. Lumsdaine. 1998. Macroeconomic news and bond market volatility. *Journal of Financial Economics* 47: 315–37. [CrossRef]
Kang, Johnny, and Carolin E. Pflueger. 2015. Inflation risk in corporate bonds. *The Journal of Finance* 70: 115–62. [CrossRef]
Kumar, Alok, and Charles M. C. Lee. 2006. Retail investor sentiment and return comovements. *The Journal of Finance* 61: 2451–86. [CrossRef]
Kyle, Albert S. 1985. Continuous Auctions and Insider Trading. *Econometrica: Journal of the Econometric Society* 53: 1315–35. [CrossRef]
Lama, Achal, Girish K. Jha, Ranjit K. Paul, and Bishal Gurung. 2015. Modelling and Forecasting of Price Volatility: An Application of GARCH and EGARCH Models. *Agricultural Economics Research Review* 28: 73–82. [CrossRef]
Lee, Byung-Joo. 2019. Asian financial market integration and the role of Chinese financial market. *International Review of Economics & Finance* 59: 490–99. [CrossRef]
Lee, Wayne Y., Christine X. Jiang, and Daniel C. Indro. 2002. Stock market volatility, excess returns, and the role of investor sentiment. *Journal of Banking & Finance* 26: 2277–99. [CrossRef]
Li, Li, and Robert F. Engle. 1998. *Macroeconomic Announcements and Volatility of Treasury Futures*. UCSD Economics Discussion Paper 98-27. Available online: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=145828 (accessed on 1 October 2025).
Liu, Feng, Deli Kong, Zilong Xiao, Xiaohui Zhang, Aimin Zhou, and Jiayin Qi. 2022. Effect of economic policies on the stock and bond market under the impact of COVID-19. *Journal of Safety Science and Resilience* 3: 24–38. [CrossRef]
Liu, Hung-Chun, and Jui-Cheng Hung. 2010. Forecasting S&P-100 stock index volatility: The role of volatility asymmetry and distributional assumption in GARCH models. *Expert Systems with Applications* 37: 4928–34. [CrossRef]
López-Cabarcos, M. Ángeles, Ada M. Pérez-Pico, Juan Piñeiro-Chousa, and Aleksandar Šević. 2021. Bitcoin volatility, stock market and investor sentiment. Are they connected? *Finance Research Letters* 38: 101399. [CrossRef]

Lu, Chia-Wu, Tsung-Kang Chen, and Hsien-Hsing Liao. 2010. Information uncertainty, information asymmetry and corporate bond yield spreads. *Journal of Banking & Finance* 34: 2265–79. [CrossRef]
Mahmood, Farrukh, and Saud Ahmed Khan. 2020. Multi-modality in the likelihood function of GARCH model. *Review of Pacific Basin Financial Markets and Policies* 23: 2050018. [CrossRef]
Mukherjee, Kedar Nath. 2019. Demystifying Yield Spread on Corporate Bonds Trades in India. *Asia-Pacific Financial Markets* 26: 253–84. [CrossRef]
Nayak, Subhankar. 2010. Investor sentiment and corporate bond yield spreads. *Review of Behavioural Finance* 2: 59–80. [CrossRef]
Nelson, Daniel B. 1991. Conditional Heteroskedasticity in Asset Returns: A New Approach. *Econometrica: Journal of the Econometric Society* 59: 347–70. [CrossRef]
Opschoor, Anne, Pawel Janus, André Lucas, and Dick Van Dijk. 2018. New HEAVY Models for Fat-Tailed Realized Covariances and Returns. *Journal of Business & Economic Statistics* 36: 643–57. [CrossRef]
Park, Beum-Jo. 2002. An outlier robust GARCH model and forecasting volatility of exchange rate returns. *Journal of Forecasting* 21: 381–93. [CrossRef]
Pham, Linh, and Oguzhan Cepni. 2022. Extreme directional spillovers between investor attention and green bond markets. *International Review of Economics & Finance* 80: 186–210. [CrossRef]
Phillips, Peter C.B., and Pierre Perron. 1988. Testing for a unit root in time series regression. *Biometrika* 75: 335–46. [CrossRef]
Piazzesi, Monika. 2005. Bond yields and the federal reserve. *Journal of Political Economy* 113: 311–44. [CrossRef]
Piñeiro-Chousa, Juan, M. Ángeles López-Cabarcos, and Aleksandar Šević. 2022. Green bond market and Sentiment: Is there a switching Behaviour? *Journal of Business Research* 141: 520–27. [CrossRef]
Rath, Prabhas Kumar. 2023. Nexus Between Indian Financial Markets and Macro-economic Shocks: A VAR Approach. *Asia-Pacific Financial Markets* 30: 131–64. [CrossRef]
Reilly, Frank K., David J. Wright, and Kam C. Chan. 2000. Bond Market Volatility Compared to Stock Market Volatility. *Journal of Portfolio Management* 27: 82. [CrossRef]
Shittu, Olanrewaju Ismail, and M. J. Asemota. 2009. Comparison of criteria for estimating the order of autoregressive process: A Monte Carlo approach. *European Journal of Scientific Research* 30: 409–16.
Spyrou, Spyros. 2013. Investor sentiment and yield spread determinants: Evidence from European markets. *Journal of Economic Studies* 40: 739–62. [CrossRef]
Tan, Dijun, and Yixiang Tian. 2009. The role of asymmetry: Evidence from Chinese Treasury bond market. *Statistics and Its Interface* 2: 57–69. [CrossRef]
Turkmen Muldur, Gozde, Serkan Yılmaz Kandir, and Yıldırım Beyazıt Onal. 2019. Investor sentiment and speculative bond yield spreads. *Foundations of Management* 11: 177–86. [CrossRef]
Verma, Rahul, and Priti Verma. 2007. Noise trading and stock market volatility. *Journal of Multinational Financial Management* 17: 231–43. [CrossRef]
Villar-Rubio, Elena, María-Dolores Huete-Morales, and Federico Galán-Valdivieso. 2023. Using EGARCH models to predict volatility in unconsolidated financial markets: The case of European carbon allowances. *Journal of Environmental Studies and Sciences* 13: 500–9. [CrossRef]
Wang, Honglin. 2023. Research on the Corporate Bond Risk Factors. *BCP Business & Management* 44: 577–83. [CrossRef]
Yong, Jordan Ngu Chuan, Sayyed Mahdi Ziaei, and Kenneth R. Szulczyk. 2021. The impact of COVID-19 pandemic on stock market return volatility: Evidence from Malaysia and Singapore. *Asian Economic and Financial Review* 11: 191. [CrossRef]
Yu, Jianfeng, and Yu Yuan. 2011. Investor sentiment and the mean-variance relation. *Journal of Financial Economics* 100: 367–81. [CrossRef]
Yung, Kenneth, and Nadia Nafar. 2017. Investor attention and the expected returns of reits. *International Review of Economics & Finance* 48: 423–39. [CrossRef]

risks

MDPI

Article

Using Daily Stock Returns to Estimate the Unconditional and Conditional Variances of Lower-Frequency Stock Returns

Chris Kirby

Department of Finance, University of North Carolina at Charlotte, 9201 University City Blvd., Charlotte, NC 28223, USA; ckirby10@uncc.edu

Abstract

If intraday price data are unavailable, then using daily returns to construct realized measures of the variances of lower-frequency returns is a natural substitute for using high-frequency returns in this context. Notably, a suitable application of this approach yields realized measures that are unbiased estimators of the unconditional and conditional variances of holding period returns for any investment horizon. I use a long sample of daily S&P 500 index returns to investigate the merits of constructing realized measures in this fashion. First, I conduct a Monte Carlo study using a data generating process that reproduces the key dynamic properties of index returns. The results of the study suggest that using realized measures constructed from daily returns to estimate the conditional and unconditional variances of lower-frequency returns should lead to substantial increases in efficiency. Next, I fit a multiplicative error model to the realized measures for weekly and monthly index returns to obtain out-of-sample forecasts of their conditional variances. Using the forecasts produced by a generalized autoregressive conditional heteroskedasticity model as a benchmark, I find that the forecasts produced by the multiplicative error model always generate lower mean absolute errors. Furthermore, the improvements in forecasting performance are statistically significant in most cases.

Keywords: forecasts; GARCH; realized volatility; realized kernel; multiplicative errors

1. Introduction

Realized variances are ex-post measures of return variation that are typically constructed by summing the squared values of high-frequency log returns (see, e.g., Andersen and Bollerslev 1998; Andersen et al. 2003; Barndorff-Nielsen et al. 2008). The use of realized variances has revolutionized methods of modeling and forecasting volatility over the past two decades. Indeed, realized variances are now employed in a wide variety of intriguing applications. Some recent examples include forecasting volatility for stocks included the S&P 500 index via machine learning methods (Zhu et al. 2023), forecasting volatility for international real estate investment trusts (Bonato et al. 2022), modeling time-varying conditional skewness in equity markets (Kirby 2024), pricing options on the Chicago Board Options Exchange volatility index (Tong and Huang 2021), developing dynamic tail-risk models to aid in measuring and managing financial risk (Chen et al. 2023), and studying volatility spillovers across cryptocurrency markets (Ben Ameur et al. 2024).

Many of the econometric studies that employ realized variances are conducted using either daily log returns or daily simple returns (see, e.g., Gorgi et al. 2019; Hansen et al. 2012, 2024; Noureldin 2022; Noureldin et al. 2012). Researchers seldom feel the need to differentiate between simple returns and log returns in such studies because doing so is

unnecessary from an empirical perspective. If the holding period for a stock or stock index is a single day, then the difference between the variance of a simple return and the variance of the corresponding log return will typically be negligible. However, the differences between the statistical properties of simple returns and those of log returns become more pronounced as the holding period increases. Thus, they are unlikely to be negligible for research that addresses asset pricing, portfolio optimization, and related topics, which is usually conducted using simple returns for weekly, monthly, or quarterly holding periods (see, e.g., Avramov et al. 2006; Kirby and Ostdiek 2012; Yogo 2006).

For instance, the covariance matrix of simple returns plays a central role in Markowitz (1952) portfolio selection. Although it would be straightforward to use realized variances computed from log returns to construct an estimator of the covariance matrix of simple returns, the estimator would typically be biased given that simple returns are nonlinearly related to log returns. Consider the case in which log returns are normally distributed. Because simple returns have a lognormal distribution under these circumstances, the variance of simple returns is typically higher than the variance of log returns in this case. This clearly suggests that it would be useful to develop a procedure for constructing realized variances that are unbiased estimators of the variances of simple returns.

More broadly, it is important to note that the high-frequency data needed to construct daily realized variances may not be available for the full sample period of interest. The first year of the Trade and Quote data provided by the New York Stock Exchange is 1993. In contrast, the coverage of the daily stock file of the Center for Research in Security Prices begins in 1926. The widespread availability of daily historical data makes it well suited for estimating the unconditional and conditional variances of lower-frequency stock returns. I investigate this approach using a new technique for constructing realized measures. Unlike the conventional construction technique pioneered by Andersen and Bollerslev (1998), the new technique delivers realized measures that are unbiased estimators of the unconditional and conditional variances of simple returns in a discrete time setting under relatively mild assumptions that are frequently invoked in the volatility modeling literature.

I begin by conducting a Monte Carlo study of the relative estimation errors that result from using the new and conventional realized measures as estimators of the unconditional and conditional variances of simple returns and log returns for a range of different holding periods. The results of the study demonstrate that my technique for constructing realized measures of the variances of simple returns works as intended. I find no evidence of bias for any holding period and the proposed realized measures deliver improvements in estimation efficiency that are comparable to those produced by conventional realized measures of the variances of log returns.

To develop further insights, I use S&P 500 index data to investigate the performance of pseudo out-of-sample variance forecasts that are constructed using the new realized measures. The empirical analysis employs the generalized autoregressive conditional heteroskedasticity (GARCH) model of Bollerslev (1986) as a benchmark and is conducted in a manner that isolates the incremental gains from using the new realized measures for modeling purposes. First, I fit a GARCH(1,1) model to simple returns for both weekly and monthly holding periods. Next, I replace the squared demeaned simple returns in the recurrence relation for the conditional variance under the GARCH(1,1) model with the corresponding realized measures. Finally, If fit the resultant specification, which is a multiplicative error model (MEM) of the type introduced by Engle (2002), to the same sample of simple returns for weekly and monthly holding periods.

Because the only difference between the recurrence relations for the conditional variances under the GARCH(1,1) and MEM(1,1) specifications is that former employs squared demeaned simple returns and the latter employs realized measures, the performance ad-

vantage of the MEM(1,1) specification (if any) is due to the incremental gains from using realized measures. I use the Giacomini and White (2006) test of equal predictive ability to assess whether the differences in performance are statistically significant. As anticipated, I find that the MEM forecasts produce smaller mean errors, smaller mean absolute errors, and smaller root mean square errors than the GARCH forecasts for every forecast horizon under consideration at both the weekly frequency and the monthly frequency. However, the results for the monthly observations are stronger from the standpoint of statistical significance. I find that the smallest t-statistic produced by the test of equal predictive ability is 2.61 in this case. Because I reject the hypothesis that the GARCH forecasts of monthly variances are just as accurate as the MEM forecasts of monthly variances at the 1% significance level, irrespective of the forecast horizon, I conclude that the proposed realized measures of the variances of simple returns deliver meaningful performance gains.

Although these results are illustrative, the proposed realized measures could obviously be exploited in other ways. For example, researchers have developed a variety of specifications that use conventional realized measures to model volatility dynamics, such as the heterogeneous autoregressive volatility (HAR) model of Corsi (2009), the high-frequency-based (HEAVY) model of Shephard and Sheppard (2010), and the realized GARCH model of Hansen et al. (2012). By replacing the realized measures constructed from high-frequency log returns with realized measures constructed from daily simple returns, any of these specifications could be used to model and forecast the conditional variances of simple returns for weekly, monthly, or quarterly holding periods.

It is also clear that the proposed realized measures can be employed to model and forecast the conditional variances of lower-frequency returns for any asset or commodity for which daily price data are readily available (e.g., the Eurodollar exchange rate, crude oil, or Bitcoin). Of course, using this approach for seasonal commodities would require a volatility model that is capable of capturing seasonality, such as a periodic MEM analog of the periodic GARCH model of Bollerslev and Ghysels (1996). But implementing this extension should be relatively straightforward from an econometric perspective.

The rest of the article is organized as follows. Section 2 shows how to construct realized measures that are unbiased estimators of the unconditional and conditional variances of simple returns. Section 3 discusses the results of the Monte Carlo study. Section 4 describes the GARCH(1,1) and MEM(1,1) specifications used to forecast conditional variances for the S&P 500 index and presents the results of the pseudo out-of-sample performance comparisons. Finally, Section 5 offers some concluding remarks.

2. Realized Measures

Suppose that $P(t_i)$ denotes the price of a stock or stock index at time t_i. Further suppose that the price is recorded at a fixed frequency such that there is always one time period between successive elements of the sequence $\{P(t_i)\}_{i=0}^{KT}$, where $K > 1$ and $T > 0$ are integers to be specified later. To develop realized measures that are unbiased estimators of the variance of simple returns in a discrete-time setting, I invoke assumptions that eliminate the need to employ the type of fill-in asymptotics that underpin the arguments of Andersen et al. (2003). Henceforth, $\mathcal{F}(t_i)$ denotes the information set that contains all prices realized prior to time t_{i+1}. I presume throughout the discussion that log returns and simple returns are weakly-stationary random variables.

2.1. Realized Measures Computed from Log Returns

Andersen and Bollerslev (1998) pioneered the use of high-frequency log returns to construct realized measures. It is easy to formulate discrete-time analogs of the basic arguments that motivate their methodology. Let $\tilde{r}(t_i, t_{i+k}) = \log P(t_{i+k}) - \log P(t_i)$ denote

the log return for the k-period interval that begins at time t_i and ends at time t_{i+k}, where $0 \leq k \leq K$. I assume for simplicity that $\mathrm{E}[\tilde{r}(t_i, t_{i+1})] = 0$ and use $\tilde{\sigma}_K^2 := \mathrm{var}(\tilde{r}(t_i, t_{i+K}))$ to denote the variance of K-period log returns.

The starting point is to consider a scenario in which the single-period log returns are serially uncorrelated. Because $\tilde{r}(t_i, t_{i+K})$ can be expressed as $\tilde{r}(t_i, t_{i+K}) = \sum_{j=1}^{K} \tilde{r}(t_{i+j-1}, t_{i+j})$, it follows immediately that

$$\tilde{\sigma}_K = \left(\mathrm{E}\left[\sum_{j=1}^{K} \tilde{r}^2(t_{i+j-1}, t_{i+j}) \right] \right)^{1/2}, \tag{1}$$

where $\tilde{r}^2(t_{i+j-1}, t_{i+j})$ denotes the square of $\tilde{r}(t_{i+j-1}, t_{i+j})$. Thus, $\tilde{v}(t_i, t_{i+K}) = (\sum_{j=1}^{K} \tilde{r}^2(t_{i+j-1}, t_{i+j}))^{1/2}$ is a realized measure of volatility that satisfies $\mathrm{E}[\tilde{v}^2(t_i, t_{i+K})] = \tilde{\sigma}_K^2$.

It is also easy to see that $T^{-1} \sum_{j=1}^{T} \tilde{v}^2(t_{(j-1)K}, t_{jK})$ and $T^{-1} \sum_{j=1}^{T} \tilde{r}^2(t_{(j-1)K}, t_{jK})$ are unbiased estimators of $\tilde{\sigma}_K^2$. But the former is a lot more efficient than the latter in general. More broadly, $\tilde{v}(t_i, t_{i+K})$ satisfies $\mathrm{E}[\tilde{v}^2(t_i, t_{i+K})|\mathcal{F}(t_i)] = \mathrm{var}(\tilde{r}(t_i, t_{i+K})|\mathcal{F}(t_i))$ under suitable assumptions about the dynamic properties of log returns. To grasp the basic requirements for conditional unbiasedness, let $\tilde{\sigma}^2(t_i, t_{i+K}) := \mathrm{var}(\tilde{r}(t_i, t_{i+K})|\mathcal{F}(t_i))$ and think about a data generating process (DGP) of the form

$$\tilde{r}(t_i, t_{i+1}) = \tilde{\sigma}(t_i, t_{i+1})\tilde{z}(t_i, t_{i+1}), \quad i = 0, 1, \ldots, KT - 1, \tag{2}$$

where $\tilde{\sigma}(t_i, t_{i+1}) \in \mathcal{F}(t_i)$, $\mathrm{E}[\tilde{z}(t_i, t_{i+1})|\mathcal{F}(t_i)] = 0$, and $\mathrm{E}[\tilde{z}^2(t_i, t_{i+1})|\mathcal{F}(t_i)] = 1$ for all i. For example, the DGP could be a GARCH(1,1) model (see Bollerslev 1986). Because a DGP of this form implies that $\mathrm{E}[\tilde{r}(t_i, t_{i+1})\tilde{r}(t_{i+j}, t_{i+j+1})|\mathcal{F}(t_i)] = 0$ for all $j \neq 0$, it follows that $\tilde{\sigma}(t_i, t_{i+K}) = (\sum_{j=1}^{K} \mathrm{E}[\tilde{r}^2(t_{i+j-1}, t_{i+j})|\mathcal{F}(t_i)])^{1/2}$ by iterated expectations.

2.2. Realized Measures Computed from Simple Returns

In a typical finance application (portfolio optimization, risk management, etc.), the analysis focuses on simple returns rather than log returns. Furthermore, the simple returns of interest are often measured at relatively low frequencies (monthly observations, quarterly observations, etc.). I therefore propose a new strategy for constructing realized measures that are unbiased estimators of the unconditional and conditional variances of simple returns. Henceforth, simple returns are just called returns.

Let $r(t_i, t_{i+k}) = P(t_{i+k})/P(t_i) - 1$ denote the return for the k-period interval that begins at time t_i and ends at time t_{i+k}. By straightforward algebra, this quantity can be expressed as

$$r(t_i, t_{i+k}) = \sum_{j=1}^{k} R(t_i, t_{i+j-1}) r(t_{i+j-1}, t_{i+j}), \tag{3}$$

where $R(t_i, t_{i+k}) = P(t_{i+k})/P(t_i)$ denotes the gross return for the k-period interval under consideration. I assume for simplicity that $\mathrm{E}[r(t_i, t_{i+1})] = 0$ and explain how to relax this assumption later on.

Now let $\sigma_K^2 := \mathrm{var}(r(t_i, t_{i+K}))$, $\sigma^2(t_i, t_{i+K}) := \mathrm{var}(r(t_i, t_{i+K})|\mathcal{F}(t_i))$, and consider a scenario in which single-period returns satisfy

$$\mathrm{cov}(r(t_{i+j-1}, t_{i+j}), R(t_i, t_{i+j-1}) r(t_{i+k-1}, t_{i+k}) R(t_i, t_{i+k-1})) = 0 \tag{4}$$

for all $j > k \geq 1$. Under these circumstances,

$$v(t_i, t_{i+K}) = \left(\sum_{j=1}^{K} R^2(t_i, t_{i+j-1}) r^2(t_{i+j-1}, t_{i+j}) \right)^{1/2} \tag{5}$$

is a realized measure of σ_K that satisfies $\mathrm{E}[v^2(t_i, t_{i+K})] = \sigma_K^2$.Furthermore, it is apparent that $v(t_i, t_{i+K})$ satisfies $\mathrm{E}[v^2(t_i, t_{i+K})|\mathcal{F}(t_i)] = \sigma^2(t_i, t_{i+K})$ under suitable assumptions about the dynamic properties of returns. This is the case, for example, if the DGP takes the form

$$r(t_i, t_{i+1}) = \sigma(t_i, t_{i+1})z(t_i, t_{i+1}), \quad i = 0, 1, \ldots, KT - 1, \tag{6}$$

where $\sigma(t_i, t_{i+1}) \in \mathcal{F}(t_i)$, $\mathrm{E}[z(t_i, t_{i+1})|\mathcal{F}(t_i)] = 0$, and $\mathrm{E}[z^2(t_i, t_{i+1})|\mathcal{F}(t_i)] = 1$ for all i. To see why, simply note that

$$\mathrm{E}[R(t_i, t_{i+j-1})r(t_{i+j-1}, t_{i+j})R(t_i, t_{i+k-1})r(t_{i+k-1}, t_{i+k})|\mathcal{F}(t_i)] = 0 \tag{7}$$

for all $j \geq 1$, $k \geq 1$, and $j \neq k$ under the DGP in Equation (6).

2.3. Some Useful Extensions

The methodology can easily be modified to address situations in which the maintained assumptions are deemed too restrictive. For instance, if single-period returns display serial correlation, then a realized kernel approach can be used to construct the realized measures. Barndorff-Nielsen et al. (2008) show that this is an effective way of addressing the presence of serial correlation that is due to microstructure effects.

To illustrate, let J denote the lag truncation value employed by the realized kernel estimator of Barndorff-Nielsen et al. (2008). An analogous estimator for simple returns can be obtained by specifying $J < K$ and computing $v(t_i, t_{i+K})$ as

$$v(t_i, t_{i+K}) = \left(\sum_{j=-J}^{J} w\left(\frac{j}{J+1}\right) g_j(P(t_i), \ldots, P(t_{i+K})) \right)^{1/2}, \tag{8}$$

where

$$g_j(P(t_i), \ldots, P(t_{i+K})) = \sum_{l=|j|+1}^{K} \frac{P(t_{i+l-1})}{P(t_i)} r(t_{i+l-1}, t_{i+l}) \frac{P(t_{i+l-|j|-1})}{P(t_i)} r(t_{i+l-|j|-1}, t_{i+l-|j|}) \tag{9}$$

and

$$w(x) = \begin{cases} 1 - 6x^2 + 6|x|^3 & 0 \leq |x| \leq 1/2 \\ 2(1 - |x|)^3 & 1/2 \leq |x| \leq 1. \end{cases} \tag{10}$$

is the weight function for the Parzen kernel.

The assumption that expected returns are equal to zero can also be relaxed. Suppose, for instance, that $\mathrm{E}[r(t_i, t_{i+1})|\mathcal{F}(t_i)] = \mu$ for all i. In this case,

$$\mathrm{E}[(1 + \mu)^{-k} R(t_i, t_{i+k}) - 1|\mathcal{F}(t_i)] = 0 \tag{11}$$

for all i and $k \geq 0$, so it is a simple matter to show that

$$\mathrm{var}\left(\frac{R(t_i, t_{i+K})}{(1+\mu)^K} \middle| \mathcal{F}(t_i) \right) = \mathrm{E}\left[\sum_{j=1}^{K} \left(\frac{R(t_i, t_{i+j-1})}{(1+\mu)^{j-1}} \right)^2 \left(\frac{r(t_{i+j-1}, t_{i+j}) - \mu}{1+\mu} \right)^2 \middle| \mathcal{F}(t_i) \right] \tag{12}$$

by mirroring the arguments for the $\mu = 0$ case. The realized measure

$$v^*(t_i, t_{i+K}) = \left(\sum_{j=1}^{K} (1+\mu)^{2(K-j)} R^2(t_i, t_{i+j-1}) (r(t_{i+j-1}, t_{i+j}) - \mu)^2 \right)^{1/2} \tag{13}$$

therefore satisfies $\mathrm{E}[v^{*2}(t_i, t_{i+K})|\mathcal{F}(t_i)] = \mathrm{var}(r(t_i, t_{i+K})|\mathcal{F}(t_i))$.

It is clear that $v^2(t_i, t_{i+K})$ is a biased estimator of $\sigma^2(t_i, t_{i+K})$ for cases in which $\mu \neq 0$, just as $\tilde{v}^2(t_i, t_{i+K})$ is a biased estimator of $\tilde{\sigma}^2(t_i, t_{i+K})$ for cases in which $\mathrm{E}[\tilde{r}(t_i, t_{i+1})|\mathcal{F}(t_i)] \neq 0$. But the results also show how to implement a simple bias correction for $v^2(t_i, t_{i+K})$. In particular, a bias-corrected realized measure for returns can be obtained by substituting a consistent estimator of $\mathrm{E}[r(t_i, t_{i+1})]$ that is available at time t_i for μ in Equation (13). The effect of this correction will typically be quite small for realized measures that are constructed from daily stock returns because the sample mean of daily returns is typically only a few basis points. This is the reason why studies that fit volatility models to daily stock returns often assume that expected returns are equal to zero (see, e.g., Visser 2011).

3. Monte Carlo Analysis

I use Monte Carlo integration to document the properties of the unbiased variance estimators discussed in Section 2. The DGP for the study is a well-known variant of the GARCH(1,1) model of Bollerslev (1986). In particular, I generate the single-period log returns from the model

$$\tilde{r}(t_i, t_{i+1}) = \kappa\tilde{\sigma}^2(t_i, t_{i+1}) + \tilde{\sigma}(t_i, t_{i+1})\tilde{z}(t_i, t_{i+1}), \tag{14}$$

$$\tilde{\sigma}^2(t_i, t_{i+1}) = \omega + \beta\tilde{\sigma}^2(t_{i-1}, t_i) + \alpha(\tilde{z}(t_{i-1}, t_i) - \gamma\tilde{\sigma}(t_{i-1}, t_i))^2, \tag{15}$$

where $\omega \geq 0$, $\beta \geq 0$, $\alpha \geq 0$, $(\beta + \alpha\gamma^2) < 1$, and $\tilde{z}(t_i, t_{i+1})$ is an NID$(0, 1)$ random variable. This specification is well suited to Monte Carlo work because it allows $\tilde{\sigma}_K^2$, σ_K^2, $\tilde{\sigma}^2(t_i, t_{i+K})$, and $\sigma^2(t_i, t_{i+K})$ to be computed analytically.[1]

Daily S&P 500 index data for the years 1946 through 2023 (19,835 observations) are used to calibrate the DGP. The data are from two sources: the daily stock file of the Center for Research in Security Prices for 3 July 1962 to 29 December 2023 and a dataset compiled by Schwert (1990) for 2 January 1946 to 2 July 1962.[2] First, I use the method of maximum likelihood to fit the model to daily log index returns subject to $\kappa = 0$ and $\omega = 0$.[3] Second, I set the values of α, β, and γ in Equations (14) and (15) equal to their maximum likelihood estimates, generate $\{\tilde{r}(t_i, t_{i+1}\}_{i=0}^{KT-1}$ with $\kappa = 0$ and $\omega = 0$, and construct $\{r(t_i, t_{i+1}\}_{i=0}^{KT-1}$ by setting $\kappa = -1/2$ and computing $r(t_i, t_{i+1}) = \exp(\kappa\tilde{\sigma}^2(t_i, t_{i+1}) + \tilde{r}(t_i, t_{i+1})) - 1$ for all i.[4] Third, I use the simulated data to calculate $\tilde{v}^2(t_i, t_{i+K})$ and $v^2(t_i, t_{i+K})$ for each $i \in \{0, K, 2K, \ldots, (T-1)K\}$. Because there are roughly 252 trading days per year for the S&P 500 index, I consider $K = 5$, $K = 21$, $K = 63$, $K = 126$, and $K = 252$ to approximate weekly, monthly, quarterly, semiannual, and annual holding periods.

Table 1 summarizes the results for 10 million simulated observations (i.e., T = 1,000,000). Panel A examines the properties of the relative estimation errors for unconditional variances. The initial six columns report the mean, mean absolute, and root mean square values of $\tilde{r}^2(t_i, t_{i+K})/\tilde{\sigma}_K^2 - 1$ and $r^2(t_i, t_{i+K})/\sigma_K^2 - 1$ for the six values of K under consideration (denoted by ME, MAE, and RMSE). For $K = 1$, the results for log returns are nearly identical to those for returns. But differences emerge as K increases.

As anticipated, the mean errors are quite small (zero to three decimal places) because $\tilde{r}^2(t_i, t_{i+K})$ and $r^2(t_i, t_{i+K})$ are unbiased estimators of $\tilde{\sigma}_K^2$ and σ_K^2. The largest RMSEs correspond to $K = 21$ for log returns and $K = 5$ for returns. An increase in the RMSE is always indicative of an increase in kurtosis, which can be expressed as 1 plus the mean square error. The smallest MAEs and RMSEs correspond to $K = 252$.

Table 1. Monte Carlo study of the relative estimation errors for unconditional and conditional variances.

Panel A												
	$\frac{\tilde{r}^2(t_i,t_{i+K})}{\tilde{\sigma}_K^2}-1$			$\frac{r^2(t_i,t_{i+K})}{\sigma_K^2}-1$			$\frac{\tilde{v}^2(t_i,t_{i+K})}{\tilde{\sigma}_K^2}-1$			$\frac{v^2(t_i,t_{i+K})}{\sigma_K^2}-1$		
K	ME	MAE	RMSE	ME	MAE	RMSE	ME	MAE	RMSE	ME	MAE	RMSE
1	−0.000	1.015	1.617	−0.000	1.015	1.618	−0.000	1.015	1.617	−0.000	1.015	1.618
5	0.000	1.027	1.692	0.000	1.025	1.666	0.000	0.615	0.864	0.000	0.612	0.856
21	0.000	1.013	1.734	0.000	1.005	1.635	0.000	0.432	0.576	0.000	0.417	0.548
63	0.000	0.997	1.698	0.000	0.984	1.532	0.000	0.317	0.412	0.000	0.284	0.362
126	0.000	0.987	1.617	0.000	0.972	1.446	0.000	0.244	0.313	0.000	0.204	0.258
252	−0.000	0.978	1.537	−0.000	0.966	1.407	0.000	0.181	0.229	0.000	0.153	0.194
Panel B												
	$\frac{\tilde{r}^2(t_i,t_{i+K})}{\tilde{\sigma}^2(t_i,t_{i+K})}-1$			$\frac{r^2(t_i,t_{i+K})}{\sigma^2(t_i,t_{i+K})}-1$			$\frac{\tilde{v}^2(t_i,t_{i+K})}{\tilde{\sigma}^2(t_i,t_{i+K})}-1$			$\frac{v^2(t_i,t_{i+K})}{\sigma^2(t_i,t_{i+K})}-1$		
K	ME	MAE	RMSE	ME	MAE	RMSE	ME	MAE	RMSE	ME	MAE	RMSE
1	−0.000	0.968	1.414	−0.000	0.968	1.415	−0.000	0.968	1.414	−0.000	0.968	1.415
5	−0.000	0.990	1.559	−0.000	0.988	1.533	0.000	0.537	0.732	0.000	0.534	0.723
21	0.000	0.995	1.676	0.000	0.988	1.579	0.000	0.380	0.506	0.000	0.362	0.476
63	0.000	0.993	1.684	0.000	0.980	1.519	0.000	0.300	0.391	0.000	0.265	0.338
126	0.000	0.985	1.613	0.000	0.971	1.441	−0.000	0.238	0.306	0.000	0.197	0.249
252	−0.000	0.978	1.536	−0.000	0.965	1.406	0.000	0.178	0.227	0.000	0.150	0.191

Note: I use Monte Carlo integration to document the performance of competing estimators of $\tilde{\sigma}_K^2 := \text{var}(\tilde{r}(t_i,t_{i+K}))$, $\sigma_K^2 := \text{var}(r(t_i,t_{i+K}))$, $\tilde{\sigma}^2(t_i,t_{i+K}) := \text{var}(\tilde{r}(t_i,t_{i+K})|\mathcal{F}(t_i))$, and $\sigma^2(t_i,t_{i+K}) := \text{var}(r(t_i,t_{i+K})|\mathcal{F}(t_i))$, where $\tilde{r}(t_i,t_{i+K})$ and $r(t_i,t_{i+K})$ denote the log return and return for the K-period interval that begins at time t_i and ends at time t_{i+K} and $\mathcal{F}(t_i) = \{\tilde{r}(t_0,t_1),\ldots,\tilde{r}(t_{i-1},t_i)\}$. The columns labeled ME, MAE, and RMSE report the mean, mean absolute, and root mean square values of the relative estimation errors for the indicated estimators. The analysis is carried out using a data generating process (DGP) of the form shown in Equations (14) and (15), where $\omega = 0$, $\beta = 0.8754$, $\alpha = 4.554 \times 10^{-6}$, $\gamma = 127.0$, and $\tilde{z}(t_i,t_{i+1}) \sim \text{NID}(0,1)$. First, I generate the sequence $\{\tilde{r}(t_i,t_{i+1}\}_{i=0}^{KT-1}$ with $\kappa = 0$ and $T = 10000000$. Next, I construct the sequence $\{r(t_i,t_{i+1}\}_{i=0}^{KT-1}$ by setting $\kappa = -1/2$ and computing $r(t_i,t_{i+1}) = \exp(\kappa\tilde{\sigma}^2(t_i,t_{i+1}) + \tilde{r}(t_i,t_{i+1})) - 1$ for all i. Finally, I calculate $\tilde{v}^2(t_i,t_{i+K}) = \sum_{j=1}^{K}\tilde{r}^2(t_{i+j-1},t_{i+j}) + 2\sum_{j=1}^{K-1}\tilde{r}(t_{i+j-1},t_{i+j})\tilde{r}(t_{i+j},t_{i+j+1})$ and $v^2(t_i,t_{i+K}) = \sum_{j=1}^{K}R^2(t_i,t_{i+j-1})r^2(t_{i+j-1},t_{i+j}) + 2\sum_{j=1}^{K-1}R(t_i,t_{i+j-1})r(t_{i+j-1},t_{i+j})R(t_i,t_{i+j})r(t_{i+j},t_{i+j+1})$ along with $\tilde{\sigma}^2(t_i,t_{i+K})$ and $\sigma^2(t_i,t_{i+K})$ for all $i \in \{0,K,2K,\ldots,(T-1)K\}$, where $R(t_i,t_{i+j}) = 1 + r(t_i,t_{i+j})$. Simple algebra yields $\tilde{\sigma}^2(t_i,t_{i+K}) = K\tilde{\sigma}_1^2 + (1-\rho)^{-1}(1-\rho^K)(\tilde{\sigma}^2(t_i,t_{i+1}) - \tilde{\sigma}_1^2)$ and $\tilde{\sigma}_K^2 = K\tilde{\sigma}_1^2$, where $\rho = \beta + \alpha\gamma^2$ and $\tilde{\sigma}_1^2 = (1-\rho)^{-1}(\omega+\alpha)$. To obtain analytic expressions for $\sigma^2(t_i,t_{i+K}) = \text{E}[R^2(t_i,t_{i+K})|\mathcal{F}(t_i)] - 1$ and $\sigma_K^2 = \text{E}[R^2(t_i,t_{i+K})] - 1$, I rely on results from the option pricing literature (see Heston and Nandi 2000, for details). Specifically, it is well known that $\text{E}[R^\tau(t_i,t_{i+K})|\mathcal{F}(t_i)] = \exp(a_K(\tau) + b_K(\tau)\tilde{\sigma}^2(t_i,t_{i+1}))$ under the DGP for the study, where $a_K(\tau)$ and $b_K(\tau)$ are given by the recurrence relations $a_K(\tau) = a_{K-1}(\tau) + \omega b_{K-1}(\tau) - \frac{1}{2}\log(1-2\alpha b_{K-1}(\tau))$ and $b_K(\tau) = \tau(\kappa+\gamma) - \frac{1}{2}\gamma^2 + \beta b_{K-1}(\tau) + \frac{(1/2)(\tau-\gamma)^2}{1-2\alpha b_{K-1}(\tau)}$ with $a_0(\tau) = b_0(\tau) = 0$. Setting $\tau = 2$ produces an expression for $\sigma^2(t_i,t_{i+K}) + 1$, which in turn yields $\sigma_K^2 + 1 = \exp(a_K(2))\text{E}[\exp(b_K(2)\tilde{\sigma}^2(t_i,t_{i+1}))]$ by the law of iterated expectations. Because $\text{E}[\exp(\xi(z+\nu)^2)] = (1-2\xi)^{-1/2}\exp(\nu^2\xi(1-2\xi)^{-1})$ given that $z \sim \text{N}(0,1)$, the law of iterated expectations also implies that $\text{E}[\exp(b_K(2)\tilde{\sigma}^2(t_i,t_{i+1}))] = \exp(\sum_{j=1}^{\infty}\omega c_{j-1} - (1/2)\log(1-2\alpha c_{j-1}))$, where c_j satisfies the recurrence relation $c_j = \beta c_{j-1} + \alpha c_{j-1}(1-2\alpha c_{j-1})^{-1}\gamma^2$ with $c_0 = b_K(2)$.

Now consider the results in the final six columns of panel A, which contain the mean, mean absolute, and root mean square values of $\tilde{v}^2(t_i,t_{i+K})/\tilde{\sigma}_K^2 - 1$ and $v^2(t_i,t_{i+K})/\sigma_K^2 - 1$ for the six values of K under consideration.[5] The realized measures show no indications of bias and are clearly much more efficient estimators of $\tilde{\sigma}_K^2$ and σ_K^2 for $K > 1$ than $\tilde{r}^2(t_i,t_{i+K})$ and $r^2(t_i,t_{i+K})$. Notice, for example, that replacing $r^2(t_i,t_{i+K})$ with $v^2(t_i,t_{i+K})$ reduces the RMSE from 1.666 to 0.856 with $K = 5$. This is a reduction of 48.6%. Furthermore, the improvements in efficiency become more pronounced as K increases. The RMSE drops from 1.407 to 0.194 for the $K = 252$ case, which is a reduction of 86.2%.

Panel B examines the properties of the relative estimation errors for the conditional variances using the same layout as panel A. Once again, the mean errors are zero to three decimal places in all cases and there are large gains in efficiency from employing the realized measures. The reduction in the RMSEs relative to those reported in panel A is an indicator of the benefits exploiting conditioning information. The RMSEs drop by 0.203 (12.6%) in all cases for $K = 1$. As K increases, the drop always becomes smaller in raw numerical terms. But the percentage drop in the RMSE does not display a monotonic relation with K. For example, the RMSE for $v^2(t_i, t_{i+K})$ drops by 0.133(15.5%) for $K = 5$.

Overall, the Monte Carlo evidence indicates that the proposed technique for constructing realized measures that are unbiased estimators of the variances of holding-period returns works as intended. It achieves improvements in efficiency that are comparable to those achieved by the conventional technique for constructing realized measures of the variances of multiperiod log returns. I now turn to an empirical application that focuses on forecasting the conditional variances of weekly and monthly S&P 500 index returns.

4. Out-of-Sample Results for the S&P 500 Index

To lay the groundwork for the discussion, assume that the objective is to forecast the variance of a financial variable $y(t+s)$ using a realization of the sequence $\{y(1), \ldots, y(t)\}$ for some $s \geq 1$. Because the GARCH(1,1) model of Bollerslev (1986) is known to perform well in a variety of settings, it is often used to construct such forecasts. If the DGP is a GARCH(1,1) specification, then $y(t+s)$ can be expressed as

$$y(t+s) = \mu + h^{1/2}(t+1,s)z(t+s), \tag{16}$$

$$h(t+1,s) = (1-\phi^{s-1})\eta + \phi^{s-1}h(t+1,1), \tag{17}$$

$$h(t+1,1) = \eta + \phi(h(t,1) - \eta) + \delta(e^2(t) - h(t,1)), \tag{18}$$

where $\mathrm{E}[z(t+s)|y(1), \ldots, y(t)] = 0$, $\mathrm{E}[z^2(t+s)|y(1), \ldots, y(t)] = 1$, and $e^2(t) = (y(t) - \mu)^2$. Thus, $h(t+1,s)$ is a conditionally-unbiased s-step-ahead forecast of $e^2(t+s)$.

Now consider an alternative s-step-ahead forecast of $e^2(t+s)$ that is constructed from a realization of the sequence $\{x(1), \ldots, x(t)\}$, where $x(t)$ is a conditionally-unbiased realized measure of the variance of $y(t)$ with dynamics that are described by an MEM specification of the type introduced by Engle (2002). If the DGP is an MEM(1,1) specification, then $x(t+s)$ can be expressed as

$$x(t+s) = m(t+1,s)u(t+s), \tag{19}$$

$$m(t+1,s) = (1-\varphi^{s-1})\varsigma + \varphi^{s-1}m(t+1,1), \tag{20}$$

$$m(t+1,1) = \varsigma + \varphi(m(t,1) - \varsigma) + \lambda(x(t) - m(t,1)), \tag{21}$$

where $u(t+s)$ is strictly non-negative and satisfies $\mathrm{E}[u(t+s)|x(1), \ldots, x(t)] = 1$. Because $m(t+1,s)$ is a conditionally unbiased s-step-ahead forecast of $x(t+s)$, it clearly has the potential to outperform $h(t+1,s)$ as a forecast of $e^2(t+s)$.

I focus on the case in which $y(t+s)$ is a weekly or monthly return on the S&P 500 index and the realized measure of its variance is constructed from daily returns. Presumably, variance forecasts based on realized measures should generally be more accurate than those based on weekly or monthly returns. I therefore use the pseudo out-of-sample forecasts produced by the GARCH(1,1) specification to benchmark the performance of the pseudo out-of-sample forecasts produced by the MEM(1,1) specification. As in Giacomini and White (2006), I conduct the analysis using limited-memory estimators of the parameters.

Because the GARCH(1,1) model implies that $e^2(t+1) = h(t+1,1)z^2(t+1)$, it is essentially a MEM(1,1) specification for $e^2(t+1)$. Furthermore, the recurrence relation for

$h(t+1,1)$ in Equation (18) can be transformed into the recurrence relation for $m(t+1,1)$ in Equation (21) by replacing $e^2(t)$ with $x(t)$ and relabeling the parameters. It is apparent, therefore, that the research design ensures that performance advantage of the MEM(1,1) specification (if any) is due to the incremental gains from using realized measures as long as the approach used to fit the GARCH and MEM specifications puts them on an equal footing. This poses no issues because fitting the GARCH specification under the assumption that $z(t+1) \sim \text{NID}(0,1)$ produces the same results as treating $e^2(t+1) = h(t+1,1)z^2(t+1)$ as an MEM specification and fitting it by assuming that $z^2(t+1)$ is a serially independent exponential random variable with a unit mean (see Engle 2002, for further elaboration).

To illustrate, suppose $W+s-1>0$ is the number of observations in a rolling window of weekly or monthly returns. For each choice of s and value of $N \in \{1,\ldots,T-W-s+1\}$, I construct an estimate of $h(t+1,s)$ for $t=N+W-1$ using the estimate of $\boldsymbol{\theta} := (\mu,\eta,\phi,\delta)$ obtained by minimizing

$$\mathcal{Q}_h(\boldsymbol{\theta};s,N) = \sum_{t=N}^{N+W-1} \frac{1}{2}\log(h(t,s)) + \frac{1}{2}\left(\frac{e^2(t+s-1)}{h(t,s)}\right) \tag{22}$$

subject to $h(N,1)=\eta$, $\mu=\hat{\mu}$, and $\eta=\hat{\eta}$, where $\hat{\mu} = W^{-1}\sum_{t=N}^{N+W-1} y(t)$ and $\hat{\eta} = W^{-1}\sum_{t=N}^{N+W-1}(y(t)-\hat{\mu})^2$. Similarly, for each choice of s and value of N, I construct an estimate of $m(t+1,s)$ for $t=N+W-1$ using the estimate of $\boldsymbol{\vartheta} := (\varsigma,\varphi,\lambda)$ obtained by minimizing

$$\mathcal{Q}_m(\boldsymbol{\vartheta};s,N) = \sum_{t=N}^{N+W-1} \log(m(t,s)) + \frac{x(t+s-1)}{m(t,s)} \tag{23}$$

subject to $m(N,1)=\varsigma$ and $\varsigma=\hat{\varsigma}$, where $\hat{\varsigma} = W^{-1}\sum_{t=N}^{N+W-1} x(t)$. The resultant estimated values of μ, $h(t+1,s)$, and $m(t+1,s)$ are denoted by $\hat{\mu}(t+1,s)$, $\hat{h}(t+1,s)$, and $\hat{m}(t+1,s)$.

Several features of this procedure are worthy of further comment. First, apart from an additive constant, $-\mathcal{Q}_h(\boldsymbol{\theta};s,N)$ and $-\mathcal{Q}_m(\boldsymbol{\vartheta};s,N)$ are the log quasi-likelihood functions that result from treating $z(t)$ as $N(0,1)$ and $u(t)$ as an exponential random variable with a rate parameter of one. Thus, the estimators of $\boldsymbol{\theta}$ and $\boldsymbol{\vartheta}$ are consistent under the usual regularity conditions for quasi-maximum likelihood estimation. Second, I use the sample mean of $y(t)$, sample variance of $y(t)$, and sample mean of $x(t)$ that are computed from the initial W observations of the rolling window as estimators of μ, η, and ς. This targeting approach simplifies optimization. Third, the procedure produces horizon-tuned forecasts because the estimates of ϕ, δ, φ, and λ are specific to the value of s under consideration.[6]

To formally compare the accuracy of $\hat{h}(t+1,s)$ and $\hat{m}(t+1,s)$ as s-step-ahead forecasts of $\hat{e}^2(t+s) = (y(t+s)-\hat{\mu}(t+1,s))^2$, I use the unconditional version of the Giacomini and White (2006) test of equal predictive ability. The test is based on the criterion

$$\Delta L(t+s) = \left|\frac{\hat{e}^2(t+s)}{\hat{h}(t+1,s)}-1\right| - \left|\frac{\hat{e}^2(t+s)}{\hat{m}(t+1,s)}-1\right|, \quad t=W,W+1,\ldots,T-s, \tag{24}$$

which is the difference between the absolute error losses produced by $\hat{h}(t+1,s)$ and $\hat{m}(t+1,s)$.[7] The null hypothesis for the test is H_0: $\text{E}[\Delta L(t+s)]=0$. Hence, inference is conducted using the t-statistic for

$$\Delta\bar{L}(s) = \frac{1}{T-W-s+1}\sum_{t=W}^{T-s}\Delta L(t+s). \tag{25}$$

If $\Delta\bar{L}(s)$ is positive and statistically significant, then the test indicates that the s-step-ahead MEM forecasts outperform the s-step-ahead GARCH(1,1) forecasts under MAE loss.[8]

The weekly and monthly index returns along with their realized variances are computed from daily index data for the years 1946 through 2023. As is typical in the finance literature, I use the actual number of trading days in a given week or given month rather than a fixed value of K for the computations. Because the daily index returns display some evidence of negative first-order serial correlation, I account for the impact of this feature by computing the realized measures as

$$v^2(t_i, t_{i+D}) = \sum_{j=1}^{D} R^2(t_i, t_{i+j-1}) r^2(t_{i+j-1}, t_{i+j}) + 2 \sum_{j=1}^{D-1} R(t_i, t_{i+j-1}) r(t_{i+j-1}, t_{i+j}) R(t_i, t_{i+j}) r(t_{i+j}, t_{i+j+1}) \quad (26)$$

rather than as shown in Section 2.[9] Here D denotes the number of trading days for the week or month in question. I specify $W = 2034$ for the weekly data and $W = 468$ for the monthly data (50% of the number of available observations in each case). To aid in interpreting the findings, I also conduct tests of equal predictive ability using weekly and monthly observations of log returns and their realized variances.

4.1. Properties of the Rolling-Window Parameter Estimates

Table 2 examines the properties of the sequence of parameter estimates produced by the rolling-window optimizations for each specification. Panels A and B present the results for weekly log returns and weekly returns. Not surprisingly, the average estimates of ϕ and φ for $s = 1$ point to strong persistence in the conditional variances for both log returns and returns. The results also indicate that the estimates of ϕ and φ are quite stable over time. In panel A, for example, the estimate of ϕ for $s = 1$ ranges from 0.943 to 0.977 and the estimate of φ for $s = 1$ ranges from 0.959 to 0.975.

The results for δ and λ in panel A display some interesting patterns. First, the average estimate of δ is somewhat smaller than the average estimate of λ for $s = 1$, $s = 3$, and $s = 6$. This finding suggests the conditional variance process of weekly log returns displays a weaker response to shocks under the GARCH specification than under the MEM specification. Second, the average estimate of δ declines monotonically with s, whereas the average estimate of λ does not. But there is a sharp drop in the average estimate of λ for $s = 12$. Although the underlying mechanism that leads to this finding is not immediately apparent, the findings for weekly returns mirror those for weekly log returns in all respects.

Panels C and D present the results for monthly log returns and monthly returns. As anticipated, the average estimates of ϕ and φ are somewhat lower than the corresponding values in panels A and B, which is consistent with returns following a stationary stochastic process. But the results still point to a substantial degree of persistence in the conditional variances. There is also more variation in the estimates of ϕ and φ over time for the monthly observations, which is an expected consequence of the sharp reduction in the number of observations in the rolling window used for estimation purposes.

Perhaps the most intriguing aspect of the results in panels C and D is that the average estimate of δ is considerably smaller than the average estimate of λ for $s = 1$, $s = 6$, and $s = 12$. This pattern suggests that the GARCH specification produces a smoother sequence of conditional variance forecasts than the MEM specification, which could indicate that the latter specification has an advantage in tracking the conditional variances. Notice that the average estimate of λ for $s = 3$ is relatively low by comparison. Because $s = 3$ for monthly observations is roughly equivalent to $s = 12$ for weekly observations, the relation between the average estimate of λ and the forecast horizon is similar at both frequencies.

Table 2. Selected properties of rolling-window parameter estimates for GARCH and MEM specifications.

Panel A: Weekly log returns												
	GARCH with s-step-ahead conditional variances						MEM with s-step-ahead conditional variances					
	ϕ			δ			φ			λ		
s	Min	Mean	Max	Min	Mean	Max	Min	Mean	Max	Min	Mean	Max
1	0.943	0.963	0.977	0.098	0.129	0.198	0.959	0.966	0.975	0.143	0.173	0.200
3	0.937	0.968	0.992	0.051	0.116	0.194	0.936	0.964	0.975	0.160	0.210	0.307
6	0.949	0.976	0.994	0.027	0.093	0.208	0.930	0.961	0.976	0.140	0.205	0.313
12	0.968	0.983	0.994	0.019	0.054	0.091	0.971	0.984	0.992	0.027	0.046	0.064
Panel B: Weekly returns												
1	0.946	0.964	0.977	0.097	0.126	0.193	0.961	0.967	0.976	0.143	0.170	0.194
3	0.941	0.968	0.991	0.056	0.116	0.191	0.938	0.965	0.976	0.159	0.208	0.304
6	0.952	0.976	0.994	0.029	0.093	0.195	0.933	0.962	0.977	0.140	0.199	0.307
12	0.968	0.983	0.994	0.020	0.054	0.091	0.972	0.984	0.992	0.028	0.048	0.069
Panel C: Monthly log returns												
1	0.874	0.940	0.975	0.056	0.087	0.132	0.821	0.867	0.922	0.472	0.552	0.667
3	0.867	0.942	0.971	0.091	0.129	0.173	0.872	0.936	0.963	0.151	0.196	0.251
6	0.861	0.934	0.964	0.084	0.141	0.175	0.882	0.929	0.960	0.212	0.449	0.812
12	0.749	0.898	0.939	0.204	0.347	0.760	0.874	0.921	0.955	0.242	0.351	0.423
Panel D: Monthly returns												
1	0.880	0.942	0.975	0.060	0.091	0.136	0.845	0.889	0.937	0.404	0.493	0.613
3	0.861	0.941	0.971	0.102	0.137	0.180	0.875	0.937	0.963	0.154	0.211	0.276
6	0.853	0.931	0.963	0.094	0.145	0.185	0.888	0.934	0.963	0.214	0.356	0.569
12	0.749	0.890	0.938	0.178	0.321	0.762	0.885	0.923	0.954	0.245	0.342	0.416

Note: The table reports selected properties of limited-memory parameter estimates that are computed using s-step-ahead forecasts of the conditional variances of weekly and monthly S&P 500 index returns. The forecasts of the conditional variances are generated by GARCH(1,1) and MEM(1,1) specifications of the form shown in Equations (16) through (18) and Equations (19) through (21). I conduct the analysis using a quasi-maximum likelihood (QML) approach that employs a rolling window of $W + s - 1$ observations to estimate the parameters. In particular, I construct the estimated values of $h(t+1, s)$ and $m(t+1, s)$ for $t = N + W - 1$ using a window of observations that begins in period N and ends in period $N + W + s - 1$, where $N \in \{1, \ldots, T - W - s + 1\}$. To compute the log quasi-likelihood functions, I treat $z(t)$ as an NID$(0, 1)$ random variable and $u(t)$ as a serially-independent exponential random variable with a rate parameter of one. Note that this methodology produces horizon-tuned forecasts of the conditional variances because the estimates of ϕ, δ, φ, and λ are specific to the value of s under consideration. I specify $W = 2034$ for the weekly data and $W = 468$ for the monthly data, which is 50% of the number of available observations in each case. The sample period is January 1946 to December 2023.

4.2. Conditional Volatility Forecasts

To develop further insights, I plot the conditional volatility forecasts for weekly returns and monthly returns. Figure 1 shows side-by-side plots of the GARCH and MEM forecasts for weekly returns. The upper panels are for $s = 1$ and lower panels are for $s = 12$. Although the side-by-side comparisons highlight the broad similarities in the forecasts for both forecast horizons, it is easy to spot a few differences. For instance, the spike in the one-step-ahead forecast of conditional volatility that follows the 1987 stock market crash is larger for the MEM specification than for the GARCH specification. But it is clear from the plots that the GARCH and MEM forecasts are highly correlated as a general rule.

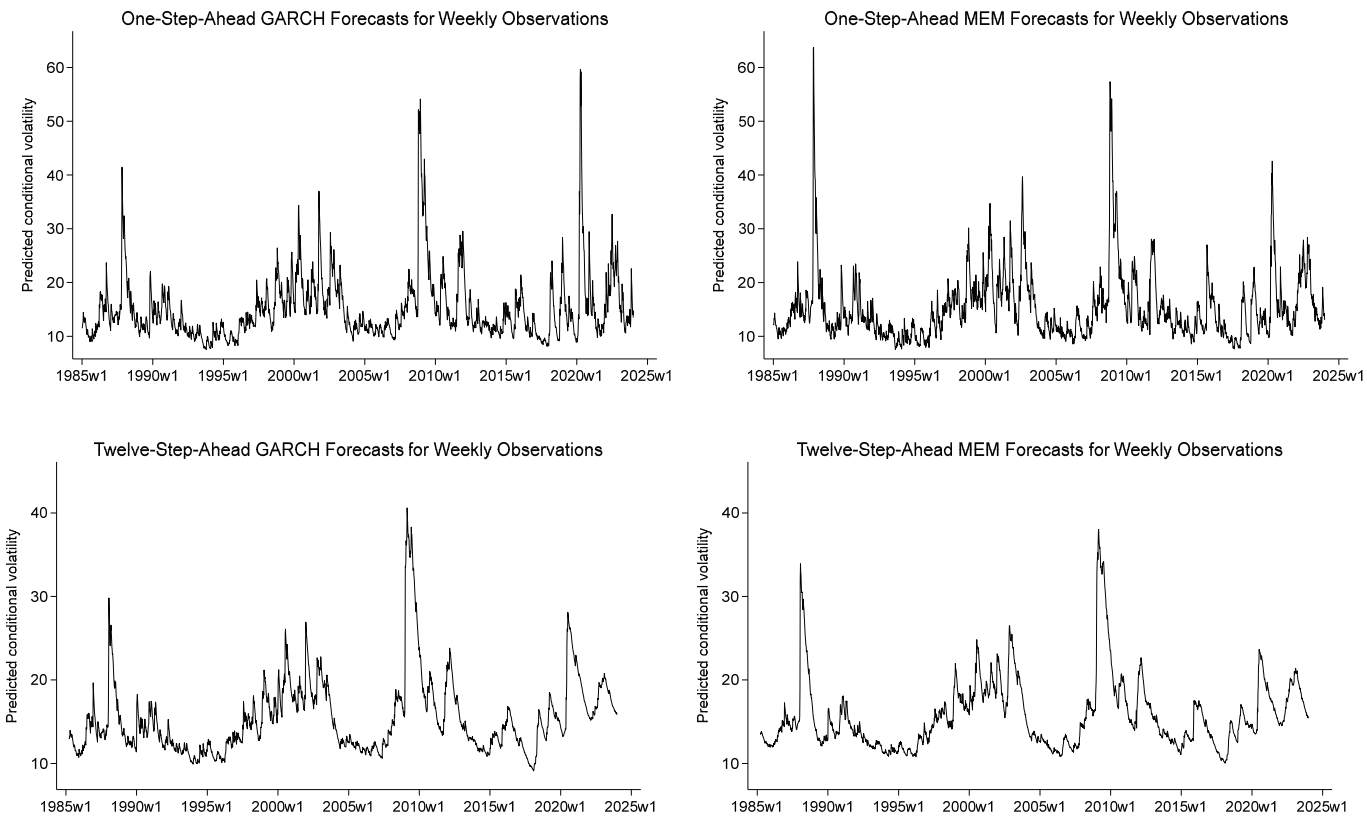


Figure 1. Out-of-sample forecasts of conditional volatility from GARCH and MEM specifications for Weekly S&P 500 Index Returns. Note: I use a rolling window of 2034 weekly observations to estimate the parameters of the GARCH and MEM specifications via quasi-maximum likelihood. The forecasts of conditional volatility, which are expressed as annualized percentage rates, are for week one of January 1985 to week four of December 2023 in the upper two panels and for week three of March 1985 to week four of December 2023 in the lower two panels. The overall sample period is January 1946 to December 2023.

Of course, this finding does not necessarily imply that the differences in the predictive ability of GARCH and MEM forecasts is negligible. If the MEM forecasts are more efficient than the GARCH forecasts, then they should have a performance advantage in formal statistical tests provided that the sample size is sufficiently large. Furthermore, the results of the Monte Carlo analysis indicate that gains from employing realized measures are inversely related to the investment horizon used for the analysis.

Consider the side-by-side plots of the GARCH and MEM forecasts for monthly returns, which are shown in Figure 2. The visual differences in the plots are certainly more pronounced in this case. Not only are the one-step-ahead GARCH forecasts relatively smooth, they are also confined to a much narrower range than one-step-ahead MEM forecasts. These features are broadly consistent with a scenario in which the realized measures are more efficient estimators of the conditional variances than the squared demeaned returns.

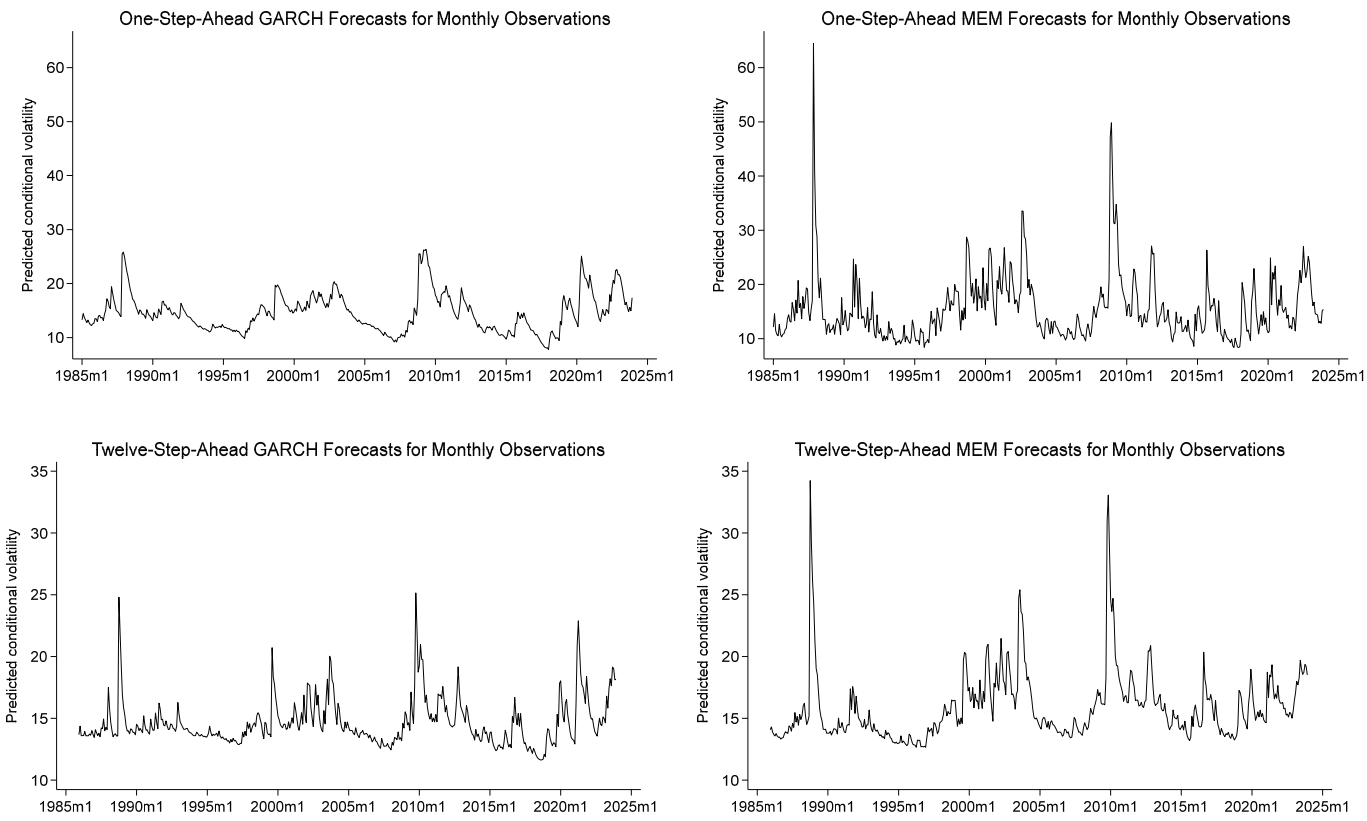


Figure 2. Out-of-sample forecasts of conditional volatility from GARCH and MEM specifications for Monthly S&P 500 Index Returns. Note: I use a rolling window of 468 monthly observations to estimate the parameters of the GARCH and MEM specifications via quasi-maximum likelihood. The forecasts of conditional volatility, which are expressed as annualized percentage rates, are for January 1985 to December 2023 in the upper two panels and for December 1985 to December 2023 in the lower two panels. The overall sample period is January 1946 to December 2023.

4.3. Hypothesis Tests

The tests of equal predictive ability provide formal evidence in this regard. The results of the tests are presented in Table 3. The initial eight columns of the table report the mean, mean absolute, root mean square, mean square values of $\hat{e}^2(t+s)/\hat{h}(t+1,s) - 1$ and $\hat{e}^2(t+s)/\hat{m}(t+1,s) - 1$ for the four choices of s: 1, 3, 6, and 12. The final three columns report $\Delta\bar{L}(s)$, its t-statistic, and the associated p-value.

The results in panel A are for weekly log returns. Notably, the MEM forecasts produce smaller MEs, MAEs, and RMSEs than the GARCH forecasts at every forecast horizon. The largest difference in the RMSE corresponds to $s = 3$: 3.604 versus 2.440. But the test of equal predictive ability produces a p-value of 0.129 in this case. Broadly speaking, however, the test favors the MEM forecasts. Note that it produces a t-statistic of 1.75 ($p = 0.079$) for $s = 6$ and 2.30 ($p = 0.021$) for $s = 12$.

The results in panel B are for weekly returns. Once again, the MEM forecasts produce smaller MEs, MAEs, and RMSEs than the GARCH forecasts at every forecast horizon. The other findings are also similar to those for weekly log returns. The test of equal predictive ability favors the MEM forecasts, yielding a t-statistic of 1.93 ($p = 0.054$) for $s = 6$ and 2.31 ($p = 0.021$) for $s = 12$.

The results in panels C and D are for monthly log returns and monthly returns. The overall pattern of the MAEs and RMSEs mirrors that in panels A and B. However, the evidence regarding the superiority of the MEM forecasts is considerably stronger at the monthly frequency. The smallest t-statistics in panels C and D are 2.36 and 2.61, which have

p-values of 0.018 and 0.009. Hence, the null hypothesis of equal predictive ability is rejected at the 1% level for every forecast horizon for monthly returns. This finding highlights the extent to which the new realized measures for lower-frequency returns deliver meaningful performance gains.

Table 3. Tests of equal predictive ability using realized measures constructed from daily observations.

Panel A: Weekly log returns											
	GARCH forecasts (s-step-ahead)				MEM forecasts (s-step-ahead)				H_0: E$[\Delta L(t+s)] = 0$		
s	ME	MAE	RMSE	MSE	ME	MAE	RMSE	MSE	$\Delta\bar{L}(s)$	t-stat	pval
1	0.067	1.083	2.172	4.716	0.039	1.061	2.127	4.522	0.022	1.41	0.157
3	0.167	1.198	3.604	12.992	0.096	1.124	2.440	5.953	0.075	1.52	0.129
6	0.211	1.262	3.802	14.459	0.170	1.216	3.481	12.118	0.046	1.75	0.079
12	0.205	1.269	3.671	13.476	0.153	1.235	3.615	13.068	0.034	2.30	0.021
Panel B: Weekly returns											
1	0.062	1.077	2.063	4.258	0.033	1.054	2.010	4.042	0.023	1.58	0.115
3	0.150	1.179	3.178	10.100	0.088	1.115	2.293	5.258	0.064	1.55	0.120
6	0.197	1.244	3.462	11.987	0.151	1.195	3.150	9.921	0.049	1.93	0.054
12	0.193	1.255	3.401	11.567	0.144	1.223	3.345	11.191	0.032	2.31	0.021
Panel C: Monthly log returns											
1	0.080	1.104	2.535	6.425	−0.059	0.971	1.900	3.612	0.133	2.83	0.005
3	0.153	1.174	2.655	7.049	−0.057	1.040	2.284	5.217	0.134	4.10	0.000
6	0.160	1.187	2.586	6.689	−0.010	1.068	2.130	4.536	0.119	2.36	0.018
12	0.157	1.195	2.705	7.316	−0.026	1.077	2.301	5.293	0.118	2.89	0.004
Panel D: Monthly returns											
1	0.064	1.080	2.177	4.740	−0.068	0.956	1.727	2.982	0.124	3.28	0.001
3	0.137	1.150	2.336	5.458	−0.052	1.029	2.028	4.114	0.121	4.04	0.000
6	0.136	1.158	2.272	5.163	−0.026	1.043	1.873	3.510	0.115	2.61	0.009
12	0.129	1.164	2.343	5.490	−0.022	1.066	2.043	4.173	0.098	3.02	0.003

Note: The table reports the results of tests of equal predictive ability for the S&P 500 index. The tests are conducted using the s-step-ahead variance forecasts produced by GARCH(1,1) and MEM(1,1) models for weekly and monthly observations (see Equations (16) through (18) and Equations (19) through (21) for details). I use a quasi-maximum likelihood approach that employs a rolling window of $W+s-1$ observations to estimate the parameters, which produces horizon-tuned forecasts because the estimates of ϕ, δ, φ, and λ are specific to the value of s under consideration. In particular, I construct the estimated values of $h(t+1,s)$ and $m(t+1,s)$ for $t = N+W-1$ using a window of observations that begins in period N and ends in period $N+W+s-1$, where $N \in \{1,\ldots,T-W-s+1\}$. The estimated values of $h(t+1,s)$ and $m(t+1,s)$ are denoted by $\hat{h}(t+1,s)$ and $\hat{m}(t+1,s)$. I base the tests on the criterion $\Delta L(t+s) = |\frac{\hat{e}^2(t+s)}{\hat{h}(t+1,s)} - 1| - |\frac{\hat{e}^2(t+s)}{\hat{m}(t+1,s)} - 1|$, $t = W, W+1, \ldots, T-s$, where $\hat{e}^2(t+s) = (y(t+s) - \hat{\mu}(t+1,s))^2$ and $\hat{\mu}(t+1,s)$ is the sample mean of $\{r(t-W+1),\ldots,r(t)\}$. The null hypothesis is H_0: E$[\Delta L(t+s)] = 0$. Hence, inference is conducted using the t-statistic for $\Delta\bar{L}(s) = \frac{1}{T-W-s+1}\sum_{t=W}^{T-s}\Delta L(t+s)$. If $\Delta\bar{L}(s)$ is positive and statistically significant, then the test indicates that the s-step-ahead MEM forecasts outperform the s-step-ahead GARCH(1,1) forecasts under the specified loss function. The initial eight columns report the mean, mean absolute, root mean square, and mean square values of $\hat{e}^2(t+s)/\hat{h}(t+1,s) - 1$ and $\hat{e}^2(t+s)/\hat{m}(t+1,s) - 1$ for the four choices of s: 1, 3, 6, and 12. I specify $W = 2034$ for the weekly data and $W = 468$ for the monthly data, which is 50% of the number of available observations in each case. The sample period is January 1946 to December 2023.

4.4. Broader Implications of the Analysis

The implications of the results for the MEM(1,1) specification extend beyond the specification itself because they suggest that replacing squared demeaned returns with the proposed realized measures can be adopted as a general strategy for improving the performance of volatility models for lower-frequency returns. Indeed, if the objective is to model the conditional variances of lower-frequency returns, then the proposed realized

measures could be in place of conventional realized measures in any existing specification that uses conventional realized measures. Some prominent examples of such specifications include the heterogeneous HAR model of Corsi (2009), the HEAVY model of Shephard and Sheppard (2010), and the realized GARCH model of Hansen et al. (2012).

As for potential applications of the methodology, it can be implemented for any asset or commodity for which daily price data are readily available. This includes international equity indexes, exchange rates, commodities, and cryptocurrencies. It should therefore prove useful in many types of research, especially if the focus is on volatility modeling, asset pricing, or risk management over medium- to long-term horizons. Depending on the application, it might be necessary to address additional features of the DGP. One that immediately comes to mind is deterministic patterns in the volatility of seasonal commodity returns. In this case, the methodology could be implemented using a volatility model that is capable of capturing seasonality, such as a periodic MEM analog of the periodic GARCH model of Bollerslev and Ghysels (1996).

4.5. Caveats

The discussion in Section 2.3 addresses two of the key assumptions that underpin the methodology and outlines techniques for relaxing these assumptions should they fail to hold in the setting of interest. But it is worthwhile to mention a few other caveats about implementing the methodology using a given specification of the DGP. Although the basic MEM(1,1) specification is useful for illustrating the incremental improvements in forecasting performance that result from using the proposed realized measures in place of squared demeaned returns, it should clearly be subject to specification tests before being adopted in a particular application. A potential concern is that, like the benchmark GARCH(1,1) specification, the basic MEM(1,1) specification is incapable of capturing leverage effects. This concern could easily be addressed by replacing Equation (21) with

$$m(t+1,1) = \varsigma + \varphi(m(t,1) - \varsigma) + (\lambda_1 I(y(t) \leq 0) + \lambda_2 I(y(t) > 0))(x(t) - m(t,1)), \quad (27)$$

where $I(\cdot)$ is the indicator function. This would yield an asymmetric MEM(1,1) analog of the threshold GARCH(1,1) specification of Glosten et al. (1993).

It is also important to consider the potential impact of phenomena, such as structural breaks, that would invalidate the assumption that simple returns are weakly stationary. Note, however, that this caveat is applicable to any volatility model that assumes weak stationarity. If an MEM(1,1) model is misspecified due the presence of a structural break or breaks, then the same is true of a GARCH(1,1) model.

5. Conclusions

The availability of high-frequency data on stock prices has transformed the volatility modeling literature over the past two decades. But there are still good arguments for using daily returns to estimate the volatility of longer-horizon returns, especially for sample periods that begin prior to 1993. Because the statistical properties of log returns differ from those of returns and the differences increase with the investment horizon, I show how to construct realized measures that are unbiased estimators of the unconditional and conditional variances of returns in a discrete-time setting, provided that the DGP satisfies relatively mild assumptions that are often invoked in the volatility-modeling literature. The empirical evidence indicates that using the proposed realized measures to compute out-of-sample forecasts of the variances of weekly and monthly returns on the S&P 500 index leads to significant improvements in forecast accuracy. Hence, the measures should be useful in research that addresses asset pricing, portfolio optimization, and related topics, which is typically conducted using returns for weekly, monthly, or quarterly holding periods.

For example, suppose an investor wants to estimate the conditional value at risk for a long position in an equity index over a one-month holding period. This could be accomplished in two simple steps. First, use the realized measures constructed from daily index returns to fit an MEM specification to the monthly index returns. Second, use the resultant sequence of estimated conditional volatilities to standardize the sequence of monthly index returns, find the quantile of the standardized index returns that corresponds to the desired confidence level for the value-at-risk criterion, and compute the conditional value at risk using the estimated conditional volatility for the month that follows the final month in the sample period.

The methodology could also be exploited in macro-finance applications. Suppose, for instance, that a researcher is interested in assessing the influence of macroeconomic variables, such as industrial production, on the volatility of monthly stock market returns. One approach for doing so would be to fit an autoregressive model to the logarithm of the gross monthly growth rate of industrial production, use the resultant parameter estimates to compute the estimated multiplicative shocks to the growth rate, and augment the conditional variance recursion of an MEM specification for monthly stock market returns with lagged values of the estimated growth-rate shocks. This approach would allow a long sample period to be used for the analysis because a century's worth of monthly data on U.S. industrial production and daily data on U.S. market returns are currently available.

Funding: This research received no external funding.

Data Availability Statement: The data for 2 January 1946 to 2 July 1962 are downloadable from https://www.billschwert.com/dstock.htm. The data for 3 July 1962 to 29 December 2023 are from a commercial data provider that charges a subscription fee to obtain data access. Please contact the Center for Research in Security Prices for subscription information.

Conflicts of Interest: The author declares no conflicts of interest.

Notes

1 See the note to Table 1 for specifics.

2 Downloadable from https://www.billschwert.com/dstock.htm, accessed on 16 August 2025.

3 The latter constraint is imposed because the nonnegativity restriction on ω is binding for the sample under consideration. This is a common finding for this model in the literature (see, e.g., Christoffersen et al. 2013).

4 The maximum likelihood estimates of the parameters are given in the note to Table 1. Notice that the simulated log returns and simulated returns have a population mean of zero by construction.

5 The results for $K = 1$ are identical to those in the first six columns because $\tilde{v}^2(t_i, t_{i+1}) = \tilde{r}^2(t_i, t_{i+1})$ and $v^2(t_i, t_{i+1}) = r^2(t_i, t_{i+1})$.

6 This approach to constructing multi-step-ahead variance forecasts is discussed in detail by Shephard and Sheppard (2010).

7 I use absolute error loss rather than squared error loss to mitigate the impact of the pronounced excess kurtosis of S&P 500 index returns, which substantially inflates the variance of $\hat{e}^2(t+s)$.

8 I use the Newey and West (1987) estimator with a lag length of $s-1$ to estimate the long-run variance of $\Delta\bar{L}(s)$.

9 Technically, the autocorrelation correction in Equation (26) could cause $v^2(t_i, t_{i+D})$ to be negative. But this never occurs in the empirical application. Under the realized kernel approach with $J = 1$, the second summation would be multiplied by $1/2$ rather than by 2.

References

Andersen, Torben, and Tim Bollerslev. 1998. Answering the skeptics: Yes, standard volatility models do provide accurate forecasts. *International Economic Review* 39: 885–905. [CrossRef]

Andersen, Torben, Tim Bollerslev, Francis Diebold, and Paul Labys. 2003. Modeling and forecasting realized volatility. *Econometrica* 71: 579–625. [CrossRef]

Avramov, Doron, Tarun Chordia, and Amit Goyal. 2006. Liquidity and autocorrelations in individual stock returns. *The Journal of Finance* 61: 2365–94. [CrossRef]

Barndorff-Nielsen, Ole, Peter Hansen, Asger Lunde, and Neil Shephard. 2008. Designing realized kernels to measure the ex-post variation of equity prices in the presence of noise. *Econometrica* 76: 1481–536.

Ben Ameur, Hachmi, Zied Ftiti, and Waël Louhichi. 2024. Interconnectedness of cryptocurrency markets: An intraday analysis of volatility spillovers based on realized volatility decomposition. *Annals of Operation Research* 341: 757–79.

Bollerslev, Tim. 1986. Generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics* 31: 307–27. [CrossRef]

Bollerslev, Tim, and Eric Ghysels. 1996. Periodic autoregressive conditional heteroscedasticity. *Journal of Business and Economic Statistics* 14: 139–51. [CrossRef]

Bonato, Matteo, Oguzhan Cepni, Rangan Gupta, and Christian Pierdzioch. 2022. Forecasting realized volatility of international REITs: The role of realized skewness and realized kurtosis. *Journal of Forecasting* 41: 303–15. [CrossRef]

Chen, Cathy, Hsiao-Yun Hsu, and Toshiaki Watanabe. 2023. Tail risk forecasting of realized volatility CAViaR models. *Finance Research Letters* 51: 103326. [CrossRef]

Christoffersen, Peter, Steven Heston, and Kris Jacobs. 2013. Capturing option anomalies with a variance-dependent pricing kernel. *Review of Financial Studies* 26: 1963–2006. [CrossRef]

Corsi, Fulvio. 2009. A simple approximate long-memory model of realized volatility. *Journal of Financial Econometrics* 7: 174–96. [CrossRef]

Engle, Robert. 2002. New frontiers for ARCH models. *Journal of Applied Econometrics* 17: 425–46. [CrossRef]

Giacomini, Raffaella, and Halbert White. 2006. Tests of conditional predictive ability. *Econometrica* 74: 1545–78. [CrossRef]

Glosten, Lawrence, Ravi Jagannathan, and David Runkle. 1993. On the relation between the expected value and the volatility of the nominal excess return on stocks. *Journal of Finance* 48: 1779–801. [CrossRef]

Gorgi, Paolo, Peter Hansen, Patryk Janus, and Siem Jan Koopman. 2019. Realized Wishart-GARCH: A score-driven multi-asset volatility model. *Journal of Financial Econometrics* 17: 1–32. [CrossRef]

Hansen, Peter, Zhuo Huang, and Howard Shek. 2012. Realized GARCH: A joint model for returns and realized measures of volatility. *Journal of Applied Econometrics* 27: 877–906. [CrossRef]

Hansen, Peter, Zhuo Huang, Chen Tong, and Tianyi Wang. 2024. Realized GARCH, CBOE VIX, and the volatility risk premium. *Journal of Financial Econometrics* 22: 187–223. [CrossRef]

Heston, Steven, and Saikat Nandi. 2000. A closed-form GARCH option valuation model. *The Review of Financial Studies* 13: 585–625. [CrossRef]

Kirby, Chris. 2024. Volatility shocks, leverage effects, and time-varying conditional skewness. *Journal of Financial Econometrics* 22: 1714–58. [CrossRef]

Kirby, Chris, and Barbara Ostdiek. 2012. It's all in the timing: Simple active portfolio strategies that outperform naive diversification. *Journal of Financial and Quantitative Analysis* 47: 437–67. [CrossRef]

Markowitz, Harry. 1952. Portfolio selection. *Journal of Finance* 7: 77–91.

Newey, Whitney, and Kenneth West. 1987. A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix. *Econometrica* 55: 703–708. [CrossRef]

Noureldin, Diaa. 2022. Volatility prediction using a realized-measure-based component model. *Journal of Financial Econometrics* 20: 76–104. [CrossRef]

Noureldin, Diaa, Neil Shephard, and Kevin Sheppard. 2012. Multivariate high-frequency-based volatility (HEAVY) models. *Journal of Applied Econometrics* 27: 907–33. [CrossRef]

Schwert, G. William. 1990. Indexes of U.S. stock prices from 1802 to 1987. *The Journal of Business* 63: 399–426. [CrossRef]

Shephard, Neil, and Kevin Sheppard. 2010. Realising the future: Forecasting with high-frequency-based volatility (HEAVY) models. *Journal of Applied Econometrics* 25: 197–231. [CrossRef]

Tong, Chen, and Zhuo Huang. 2021. Pricing VIX options with realized volatility. *Journal of Futures Markets* 41: 1180–200. [CrossRef]

Visser, Marcel. 2011. GARCH parameter estimation using high-frequency data. *Journal of Financial Econometrics* 9: 162–97.

Yogo, Motohiro. 2006. A consumption-based explanation of expected stock returns. *The Journal of Finance* 61: 539–80. [CrossRef]

Zhu, Haibin, Lu Bai, Lidan He, and Zhi Liu. 2023. Forecasting realized volatility with machine learning: Panel data perspective. *Journal of Empirical Finance* 73: 251–71. [CrossRef]

risks

MDPI

Article

Historical Perspectives in Volatility Forecasting Methods with Machine Learning

Zhiang Qiu [1], Clemens Kownatzki [2], Fabien Scalzo [1] and Eun Sang Cha [1,3,*]

[1] Seaver College, Pepperdine University, Malibu, CA 90263, USA; zhiang.qiu@pepperdine.edu (Z.Q.); fabien.scalzo@pepperdine.edu (F.S.)
[2] Pepperdine Graziadio Business School, Pepperdine University, Malibu, CA 90263, USA; clemens.kownatzki@pepperdine.edu
[3] Institute of Advanced Machinery, Design & Technology, Korea University, Seoul 136-713, Republic of Korea
* Correspondence: eunsang.cha@pepperdine.edu

Abstract: Volatility forecasting for financial institutions plays a pivotal role across a wide range of domains, such as risk management, option pricing, and market making. For instance, banks can incorporate volatility forecasts into stress testing frameworks to ensure they are holding sufficient capital during extreme market conditions. However, volatility forecasting is challenging because volatility can only be estimated, and different factors influence volatility, ranging from macroeconomic indicators to investor sentiments. While recent works show promising advances in machine learning and artificial intelligence for volatility forecasting, a comprehensive assessment of current statistical and learning-based methods is lacking. Thus, this paper aims to provide a comprehensive survey of the historical evolution of volatility forecasting with a comparative benchmark of key landmark models, such as implied volatility, GARCH, LSTM, and Transformer. We open-source our benchmark code to further research in learning-based methods for volatility forecasting.

Keywords: volatility forecasting; risk management; deep learning; time series analysis; GARCH; LSTM; transformer

1. Introduction

Risk management is essential to financial institutions because it is required by regulators and highly relevant to their performance. An integral aspect of this risk stems from the movement of the equity market, especially the market's volatility. Volatility can be used to determine the risk exposure of a portfolio (R. Engle 2004), the anticipated fluctuations throughout the duration of an option (Black and Scholes 1973), and the bid–ask spread of options as well as their underlying asset (Bollerslev and Melvin 1994). A model that helps financial institutions forecast the volatility of their holdings would provide a clearer picture of their risk and facilitate the process of risk management and decision-making.

Volatility cannot be observed; therefore, it has to be estimated, which is why we embark on this journey to survey different volatility models. There are various ways that researchers have used to estimate volatility; the two most common ones are realized volatility, which is the annualized standard deviation of log returns, and implied volatility, which is a forward-looking metric calculated from options. While the Black–Scholes model as the basis of implied volatility has unrealistic assumptions, it is based on actual market data and has been a standardized way of quantifying volatility. In this paper, we choose realized volatility as our prediction target and use implied volatility as one of the ways to predict it. Other than implied volatility, researchers have employed a wide

range of other models, ranging from traditional Generalized Autoregressive Conditional Heteroskedasticity (GARCH) (Bollerslev 1986) to Neural Network frameworks such as Long Short Term Memory (LSTM) (Hochreiter and Schmidhuber 1997) and transformers (Vaswani et al. 2023). In this paper, we compare the relative performance of these models, showing how machine learning models have significantly outperformed earlier ones, while also discussing the limitations of these models, some of which are hard to interpret intuitively.

Volatility exhibits a series of stylized facts, including temporal clustering (Mandelbrot 1997; Kim and Shin 2023), long memory (Poon and Granger 2003), heavy tails (Cont 2001), leverage effect (McAleer and Medeiros 2008; Aıt-Sahalia 2017; Engle and Ng 1993; Tversky and Kahneman 1991; Christie 1982; Bekaert and Wu 2000), and mean reversion (Goudarzi 2013). These stylized attributes make volatility forecasting possible and are detailed in Appendix A. Nevertheless, predicting volatility remains a challenging endeavor. Volatility is influenced by a wide range of factors, including macroeconomic phenomena, corporate earnings reports, interest rates, global commodity price trends, and psychology (Shiller 1999). The interactions among these factors can be complex. Volatility as a proxy for risk is easily mistaken for uncertainty (Knight 1921). Volatility can be estimated when one knows the range of possible outcomes and can thus assign a probability distribution to these outcomes—in other words, measurable unknowns. Uncertainty, by contrast, refers to unknown unknowns, and, hence, no probability distribution can be assigned to these unknown outcomes. Examples include exogenous shocks, such as geopolitical tensions, natural calamities, or abrupt regulatory shifts, which can yield immediate and pronounced volatility spikes and are inherently challenging to anticipate.

The benefits and challenges of forecasting volatility have garnered attention from many, and ongoing revisions have been undertaken to include the latest developments in this domain (Andersen et al. 2005). There are numerous studies surveying volatility forecasting methods (Ge et al. 2023; Sezer et al. 2020); however, they either focus on a specific type of model or do not incorporate the current state-of-the-art models. We are not aware of a recent comprehensive review that provides a clear explanation and compares the foundational and cutting-edge volatility models side by side. Within this paper, we survey the evolution of volatility forecasting models and evaluate their performance using a representative dataset from the Standard and Poor's 500 index (S&P 500). Section 2 provides a comprehensive literature review; Section 3 discusses the models we used and their advantages and disadvantages. Section 4 discusses how the models perform during extreme periods of market volatility, like the 2008 Financial Crisis and the 2020 COVID-19 Pandemic. Section 5 illustrates the results, and Section 6 discusses the implications of these results. Section 7 provides a conclusion, and Section 8 provides some future work directions. The contributions of this paper are as follows:

1. We survey the evolution of volatility forecasting models, transitioning from traditional AR and implied volatility models to contemporary variations of the transformer models representing the current state of the art.
2. We select a representative model from each category and conduct a systematic review to show their respective performances, paving the way for subsequent model developments.
3. We open-source our analysis framework and highlight the advantages and disadvantages inherent to each type of model.

2. Literature Review

2.1. Overview

In volatility forecasting, numerous models have been employed, and many of them have been applied in conjunction with others. We selected models that are well documented in papers. Given the extensive nature of research in this field, we focused on those related to volatility forecasting, validated by multiple datasets, and made an important paradigm shift. For organizational purposes, we have categorized these models into four primary classifications: statistical models, implied volatility, recurrent neural networks (RNNs) (Connor et al. 1994; Chung et al. 2014), and transformers, as detailed in Figure 1. In the following sections, we first examine each category and its models. Then, we discuss the advantages that led to the emergence of each and their limitations. Finally, we highlight the landmark models within these categories and perform a comparative analysis.

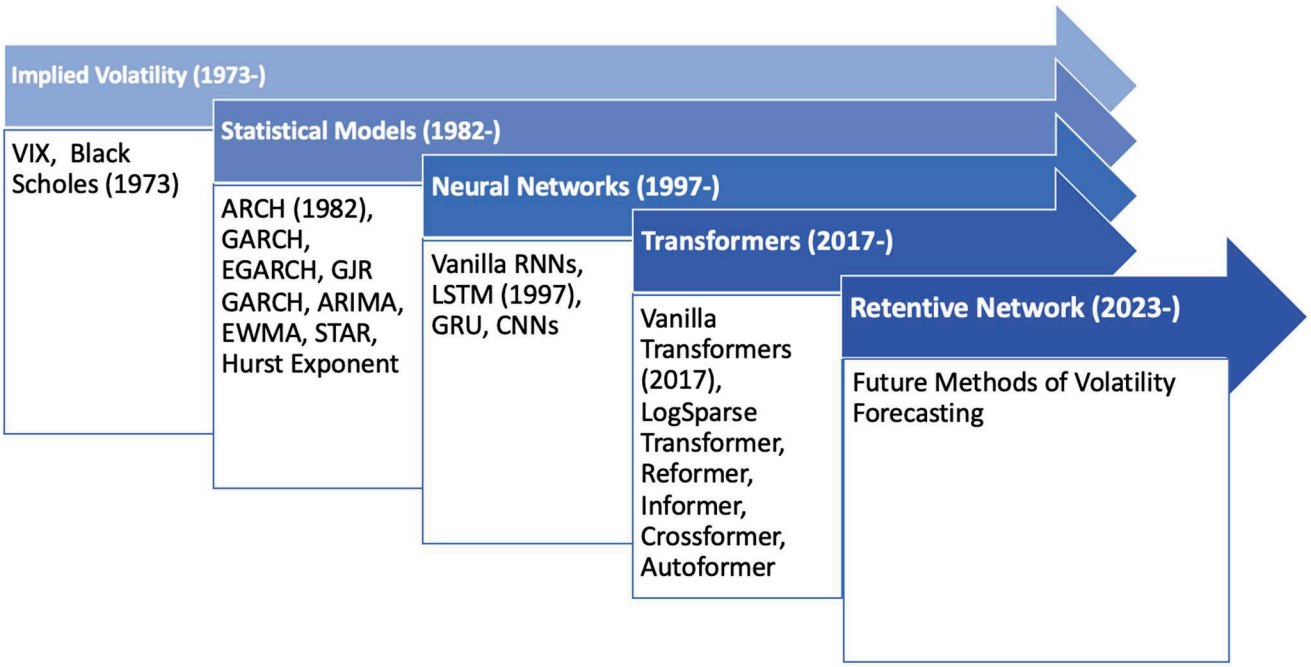


Figure 1. Timeline for the evolution of volatility prediction models.

2.2. Implied Volatility

The Chicago Board Options Exchange's CBOE Volatility Index (VIX) and implied volatility (IV) are widely used as predictors for future volatility. Often referred to as the "fear index", the VIX mirrors the market's 30-day anticipated volatility (Whaley 2009). Unlike its original derivation, which was based on a narrow set of strike prices to determine implied volatility, today, the VIX is based on the methodology of a volatility swap (Derman 1999). However, rather than being an actual swap, a volatility swap is a forward contract on the realized variance (Diamond 2012). The computation of VIX uses two months of the latest option data while interpolating between the nearest and second nearest expiration months to create a consistent 30-day window of expected volatility. Specific option strikes are then chosen for the VIX calculation, as elaborated in the VIX white paper (CBOE 2019). IV is viewed as a reliable predictor because it reflects actual investor expectations (Poon and Granger 2003). However, IV has its limitations. It can only be estimated through an iterative procedure using option prices.

2.3. Statistical Models

To address the stylized facts, Robert Engle's Autoregressive Conditional Heteroskedasticity (ARCH) model (R. F. Engle 1982) was initially adopted for volatility forecasting. This was followed by the introduction of the Generalized Autoregressive Conditional Heteroskedasticity (GARCH) model by his student Tim Bollerslev, with the common goal of leveraging the occurrence of volatility clustering (Andersen 2018). Since its inception, numerous scholars have employed and adapted the GARCH model. For example, the asymmetric GARCH (AGARCH), Exponential GARCH (EGARCH), and Glosten Jagannathan Runkle GARCH (GJR GARCH) observed that negative shocks have a more substantial impact on the variance than positive shocks. According to the survey paper Bollerslev wrote in 2008 (Bollerslev 2008), there were already numerous variants of the original ARCH model, and this number is continuously growing. Other than GARCH models, other statistical models include the simple moving average (SMA) (Johnston et al. 1999), exponentially weighted moving average (EWMA) (Holt 2004), Smooth Transition Exponential Smoothing (Taylor 2004), autoregressive integrated moving average (ARIMA) (Box and Pierce 1970), and smooth transition autoregressive (STAR) (Bildirici and Ersin 2015). These models accommodate factors such as seasonality, long-term trends, autoregressive, and moving averages. Among them, SMA creates a smooth curve from past data over a defined period, while EWMA emphasizes recent data more by giving more weight to recent observations. The Smooth Transition Exponential Smoothing model uses a logistic function based on a selected variable for its smoothing effect. ARIMA combines autoregressive techniques and moving averages, and STAR focuses on detecting nonlinear trends in data.

A common characteristic among AR models is their reliance on historical values and stationarity for predictions. For a time series to be stationary, its statistical properties, such as mean, variance, and covariance, must not be a function of time. To test whether a time series is stationary, the Augmented Dickey–Fuller (ADF) test is typically implemented. Stock prices themselves are not stationary, as they tend to increase with a drift term; however, their log returns are usually stationary. The requirement of converting price data to log return data is a significant limitation for AR models when implemented in real-time. Volatility time series generally require treatment to become stationary, and the specific treatment depends on whether the time series is deterministic. A deterministic series means the same input generates the same output every time, whereas a stochastic series produces different results. While there are deterministic patterns in volatility, such as long-term mean reversion and volatility clustering, a significant portion of volatility arises from uncertainty, as discussed in the introduction. Although chaos theory suggests an alternative explanation that volatility is deterministic yet highly sensitive to initial conditions, stochastic models like the Heston, GARCH, and SABR models are widely used. The detrending and differencing methods aim to remove the trend and seasonality from the time series and focus on the stochastic part instead. In general, detrending removes the trend from the time series, a technique that can be applied if the volatility is deterministic. Differencing, which is subtracting the previous observation from the current one, is more appropriate for stochastic cases. They were both widely adopted in statistical and machine learning models (Raudys and Goldstein 2022; Granger and Joyeux 1980). If the underlying data are non-stationary, there are alternative methods such as the Hurst Exponent, which measures whether the trend shows momentum or is mean reverting (Hurst 1951; Qian and Rasheed 2004).

2.4. Tree-Based Models and PCA

With its ability to learn from and make predictions based on data, machine learning has been applied to more and more fields, and finance is no exception. Unlike econometric

models that aim to be parsimonious by limiting the number of parameters, machine learning embraces the use of a vast number of parameters (Kelly and Xiu 2023). This approach has led to the adoption of many new techniques for volatility forecasting, either as entirely new methods or as extensions of existing ones. The following paragraphs will start with traditional machine learning (ML) models and extend to neural networks (NNs), both used extensively for volatility analysis.

Decision trees (DTs) (Loh 2011), random forests (RF) (Breiman 2001), and XGBoost (Chen and Guestrin 2016) are among the foundational applications for machine learning. Decision trees consist of a supervised learning algorithm that ascertains the value of a target variable by deducing straightforward decision rules from the data's features. The Random Forest (RF) algorithm operates as an ensemble of these decision trees, chosen through stochastic processes. XGBoost is an optimized distributed gradient boosting library rooted in decision tree algorithms (Zhang et al. 2023). Different from random forest, XGBoost uses a boosting method to combine trees together so that each tree corrects the error of the previous one.

Furthermore, another well-used machine learning model is Principal Component Analysis (PCA). PCA reduces the dimension of the dataset while ensuring the principal components are still consistent estimators of the true factors (Stock and Watson 2002). For example, Ludvigson and Ng found a volatility factor and a risk premium factor that contain significant information about future returns (Ludvigson and Ng 2007). Such analysis is further improved by assigning weights to predictors that reflect their relative forecasting strength (Huang et al. 2022).

2.5. Neural Networks

While traditional ML methods performed well in short-term forecasting, they also have several limitations. Their ability to predict long-term and complex volatility is limited, and missing values, which are not uncommon in practice, can cause significant issues for these traditional models. In order to address these problems, neural networks (NNs) were introduced (Pranav and Hegde 2021). NNs were inspired by the biological neural networks in human brains and can detect complex patterns in nonlinear form (Hornik et al. 1989). Structurally, a neural network (NN) is composed of multiple layers. Each layer features neurons that are interconnected through weighted links, which are then adjusted during the training process. This adjustment is done by using backpropagation to compute the gradient of the loss function for each weight using the chain rule.

NNs were traditionally used for tasks like image and speech recognition (Abdel-Hamid et al. 2014), targeted marketing (Venugopal and Baets 1994), and autonomous vehicles (Pomerleau 1988). When applied to time series, the NNs can recognize the relationship between past and current inputs and use them to predict future outputs. While various NNs have been applied (Chow and Leung 1996; Marcek 2018), recurrent neural networks (RNNs) are particularly prominent for time series-related tasks. RNNs use activation functions to model nonlinear relationships. They will retain information specific to a particular timestep and sequentially update it at each future step. This makes them ideal for time series analysis. In the RNN process, data are initially fed to produce a preliminary result. This result is then contrasted with the actual outcome using the loss function. Subsequently, backpropagation will be used to fine-tune the gradient for each neuron in the network. This iterative process optimizes each neuron's weight. However, despite their potential, RNNs have limitations, especially the vanishing gradient problem ("Gradient Flow in Recurrent Nets: The Difficulty of Learning LongTerm Dependencies" Kolen and Kremer 2009). In backpropagation, derivatives are calculated layer by layer from the end to the beginning. As per the chain rule, when these derivatives are successively multiplied, they can diminish

exponentially, causing them to vanish. Similarly, the gradient could also explode if the opposite happens. These lead to the RNN's failure to learn long-term dependencies. RNNs are also computationally expensive and cannot be parallelized, which makes training an RNN difficult.

Algorithmic methods such as Long Short Term Memory (LSTM) and Gated Recurrent Unit (GRU) have been introduced to address these problems. Both LSTM and GRU extend the capabilities of RNNs. These models employ gates to update or remove information to the hidden states to address the long-term dependencies. The GRU is a simplified version of the LSTM that combines and reduces parameters. While less powerful, it is relatively faster compared with LSTM. Both LSTM and GRU are widely used in time series forecasting, both on their own and in conjunction with other models (Kim and Won 2018; Ozdemir et al. 2022; Gu et al. 2020). However, the sequential nature of LSTMs and GRUs makes them time-intensive when working with larger data.

Another neural network used is Convolutional Neural Networks (CNNs), which was widely used in computer vision for image recognition tasks. Using CNNs would require image format data, which can be achieved in different ways, such as snapshots of data in a bounded period (Gudelek et al. 2017) or multiple technical indicators, each with the same length (Sezer and Ozbayoglu 2018).

2.6. Transformers

Transformers address the problems of RNNs by taking advantage of GPUs to do parallel computing. The original use of a transformer was in translating languages, but soon it saw extensive use in different tasks like audio processing (Huang et al. 2018), computer visions (Liu et al. 2021), and time series (Ahmed et al. 2022; Zerveas et al. 2021). Transformers abandon recurrence entirely and use the attention mechanisms instead. Transformers have an encoder–decoder structure. The encoder maps the input sequence and produces a continuous representation, and the decoder chooses what and how much previously encoded information to access. The encoder's attention mechanism derives attention scores from input vectors of queries, keys, and values. These scores determine the weight of each piece of information in predicting every time step. A dot product was used for simplicity in the original transformer model, whereas many different approaches were introduced in later works (Wen et al. 2023). To further refine and simplify the process, an array of attention mechanisms emerged.

Notable examples include the LogSparse Transformer (Li et al. 2019), Reformer (Kitaev et al. 2020), Informer (Zhou et al. 2021), Crossformer (Zhang and Yan 2023), and Autoformer (Wu et al. 2021). As research in this area is ongoing, these mechanisms constantly advance the state of the art. In their distinct ways, these innovative mechanisms cut time and memory requirements compared to the original transformer.

The LogSparse Transformer introduces convolutional self-attention. It generates queries and keys using causal convolution, prioritizing more recent information for immediate step forecasting. The Reformer uses locality-sensitive hashing to replace simple dot products and reversible residual layers to replace standard residuals. The Informer shares similarities with LogSparse by utilizing the sparsity found in the self-attention probability distribution. However, the Informer identifies a long-tail distribution within the attention distribution. Therefore, it leverages the fact that a small number of dot products produce the majority of attention to selectively choose only the top queries and replaces vanilla self-attention with ProbSparse self-attention. The Crossformer identifies a gap in cross-dimensional dependency modeling. To remedy this, it introduces a Two-Stage Attention (TSA) layer to bridge this deficiency. Autoformer model utilizes a decomposition layer to separate the time series into long-term trends, seasonality, and random components.

Then, it replaces the self-attention with the autocorrelation mechanism, which extracts frequency-based dependencies from queries and keys instead of the vanilla dot product.

Despite transformers being regarded as a state-of-the-art (SOTA) way of forecasting volatility, Zeng et al. (2023) challenged this notion by introducing a straightforward single-layer linear model that surpassed all current transformer models across nine datasets. They argued that transformer architectures, despite their success in NLP, may not be suitable for time series forecasting. This suggests that the self-attention mechanism is inherently anti-order. Zeng argued that while this might not significantly impact sentences, as they retain most of their meaning even if the sequence of the words is changed, it is problematic for time series where the continuous sequence order is vital.

This perspective quickly gained attention. For instance, Nie et al. introduced PatchTST (Nie et al. 2023), addressing the anti-order issue by segmenting time steps into subseries-level patches. Although Cirstea et al. (2022) first introduced the concept of patches for simplifying complexities, PatchTST was the first to utilize them as input units. Furthermore, they incorporated a channel-independence technique previously validated by Zheng et al. (2014). This ensures the input token is derived from a single channel, in contrast to earlier transformers that adopted channel-mixing methods. Their results indicated a marked improvement over both standard transformers and the linear model proposed by Zeng et al. (2023). Other than ignoring the temporal dependencies, the transformer also faces common shortcomings that many models face, such as being prone to overfitting, dependency on stationarity, and hard to interpret. Additionally, they are computationally intensive, memory-demanding, and prone to high latency when deployed in real-time applications, highlighting the need for further refinement and optimization.

While constructing the paper, Microsoft and Tsinghua University introduced a novel architecture called the Retentive Network (Sun et al. 2023). This architecture is presented as an enhancement to the Transformer model, which reduces inference costs and long-sequence memory complexity. Although its primary intention is for use in language models, similar to the evolution of transformers, this architecture may find broader applications, like time series.

2.7. Hybrid and Ensemble Models

While statistical models like EWMA and GARCH differ from neural networks, they still share many similarities that make hybrid designs possible (Kim and Won 2018). For example, the GARCH model can be viewed as an RNN without the output layer and activation functions (Zhao et al. 2024). Among the hybrid models, the most notable one is GARCH-LSTM, where GARCH's output replaces the output gate of LSTM, and the GARCH parameters are used as features in the LSTM network (García-Medina and Aguayo-Moreno 2024). Various types of GARCH LSTM models have shown performance improvement, and the GARCH model adds valuable input to the NNs. For example, Koo and Kim developed a volume-up method to avoid biases in the input distribution and improve the result. They compared their VU strategy with ten different GARCH-LSTM variants, such as EGARCH with LSTM (Koo and Kim 2022). Another hybrid model is GARCH-MIDAS (Mixed Data Sampling), which keeps the GARCH process for the short term and uses macroeconomic variables as part of the long-term component (Asgharian et al. 2013; Fang et al. 2020). Ensemble models also show performance improvement. For example, He et al. proposed an ensemble model based on CNN, LSTM, and ARMA that improved accuracy and robustness compared with individual models (He et al. 2023). Olorunnimbe and Viktor used an ensemble of temporal transformers and showed a 40% to 60% improvement over the baseline temporal transformer (Olorunnimbe and Viktor 2024).

3. Methods

3.1. Overview

We will discuss four milestone models: Generalized Autoregressive Conditional Heteroskedasticity, implied volatility, Long Short Term Memory, and transformer. Table 1 summarizes the advantages and disadvantages of each model. We selected these four models because they are representative of each of the categories we discussed and are widely used. We aim to trace the historical evolution of volatility forecasting techniques by showing how newer models compare to earlier ones. In the following section, we show the methodology of each model. We used Python 3 to implement these models and tested their performance when used to forecast volatility. We evaluate their performance based on root mean square error (RMSE), which is a widely used metric for predicting numerical data.

Table 1. Summary of the advantages and disadvantages of the different models.

Model	Advantages	Disadvantages
Implied Volatility (IV)	• Forward-looking • Based on real market expectations	• Relatively low accuracy • Sensitive to market noise, especially when trading volume is low
Statistical Models (GARCH, EGARCH, GJR-GARCH)	• Easy to Interpret • Incorporate established stylized facts • Fast to compute • Works well with all data sizes	• Relatively low accuracy • Assume data are stationary
Recurrent Neural Networks (RNNs, LSTMs, GRUs)	• Capture nonlinear patterns • High Accuracy	• Prone to overfitting • Require large datasets • Take a long time to train
Transformers	• Computation can be parallelized • Capture long-term dependencies • High Accuracy	• High computational power requirement • Prone to overfitting • Require large datasets • Take a long time to train

3.2. Data Processing

We used thirty years of daily S&P 500 data from Yahoo Finance, 1 October 1993 to 1 October 2023. We used the first twenty-eight years for training and the last two years for testing, as shown in Figure 2. The actual training data (3 December 1993 to 28 September 2021) is slightly shorter than twenty-eight years because of data loss when processing data and calculating rolling returns. We calculated the realized volatility as the annualized standard deviation of 22 rolling trading days' (as an approximation for one month in time) log return. We also tried 11 trading days (approximately 15 days), and the RMSPE for GARCH and IV are relatively stable but nearly doubled for the transformer and tripled for LSTM, at 0.10 and 0.14, respectively. The logarithmic return for a given day t is represented as follows:

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

where:

- r_t: Log return on day t.
- P_t: Price on day t.
- P_{t-1}: Price on the previous day, day $t-1$.

The average log return over 22 trading days is as follows:

$$\bar{r} = \frac{1}{22}\sum_{i=0}^{21} r_{t-i} \tag{2}$$

The realized volatility over 22 trading days is given as follows:

$$\Sigma = \sqrt{\frac{1}{21}\sum_{i=0}^{21}(r_{t-i} - \bar{r})^2} \tag{3}$$

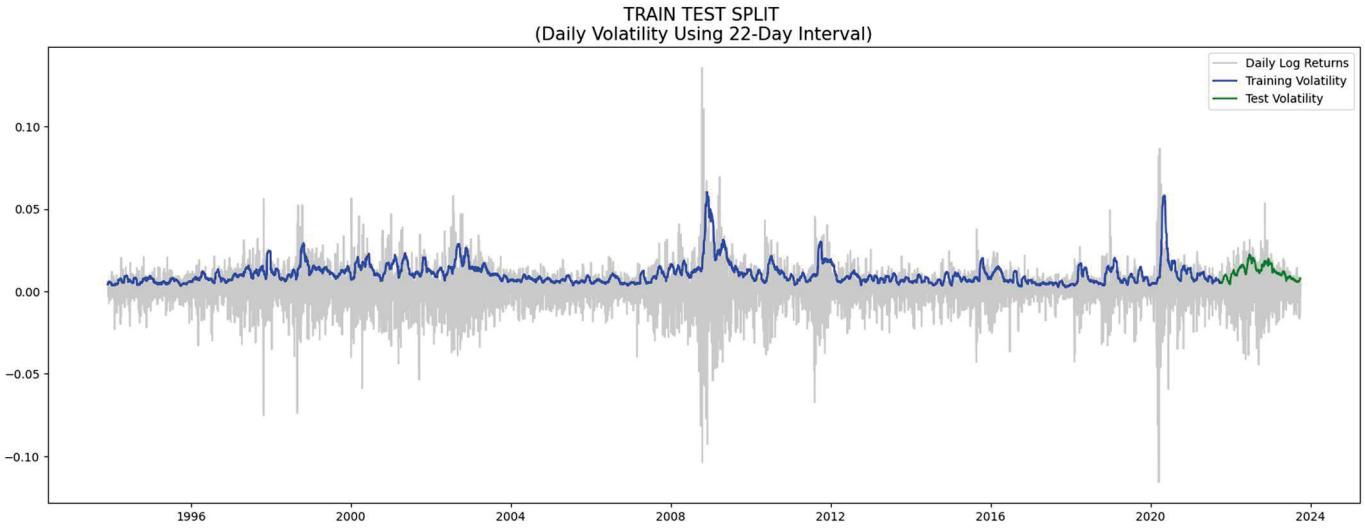


Figure 2. Training and testing data intervals.

3.3. Implied Volatility

What distinguishes IV from other models is that it is forward-looking. Implied volatility captures the market's expectation of the volatility for the next 22 trading days, calculated backward from the option's price using the Black Scholes Merton Formula:

$$c = S_0 N(d_1) - Ke^{-rT} N(d_2) \tag{4}$$

$$p = Ke^{-rT} N(-d_2) - S_0 N(-d_1) \tag{5}$$

where:

$$d_1 = \frac{\ln\left(\frac{S_0}{K}\right) + \left(r - q + \frac{\sigma^2}{2}\right)T}{\sigma\sqrt{T}} \tag{6}$$

$$d_2 = d_1 - \sigma\sqrt{T} \tag{7}$$

Explanation of terms:

- c: Price of the European call option.
- p: Price of the European put option.
- S_0: Current stock price.
- K: Strike price of the option.
- T: Time to maturity (in years).
- r: Risk-free rate.
- q: Continuous dividend yield.
- $N(.)$: The probability that a variable with a standard normal distribution will be less than x.
- σ: The standard deviation of the stock's returns.

While it is impossible to invert the function to calculate implied volatility directly as a function of other variables, an iterative approach can be used to search for implied volatility (Hull 2018). In this paper, we used the corresponding daily close of the VIX index as the implied volatility and compared it with the realized volatility of the S&P 500. The VIX data are available from the Chicago Board Options Exchange (CBOE).

3.4. GARCH

The ARCH model was introduced prior to the GARCH model for forecasting volatility. The name, Autoregressive Conditional Heteroskedasticity, means that volatility depends on the time series value in previous periods and some error terms. GARCH is a variant of the ARCH model that addresses the problem of predictions being bursty, which means the predictions vary by a huge amount day by day. This enhancement is achieved by taking the previous day's volatility into the current day's calculation, alongside the ARCH model's time series value and error term. The equation for GARCH(p, q) is as follows:

$$\sigma_t^2 = \alpha_0 + \sum_{i=1}^{p} \alpha_i \varepsilon_{t-i}^2 + \sum_{j=1}^{q} \beta_j \sigma_{t-j}^2 \tag{8}$$

where:

- σ_t^2: Conditional volatility at time t.
- α_0: positive empirical parameters.
- α_i: Non-negative empirical parameters.
- ε_{t-i}^2: the squared residual at time $t - i$.
- β_j: Non-negative empirical parameters.
- σ_{t-j}^2: Variance of the return series at time $t - j$.

While there are many different versions of GARCH models, we used a simple GARCH (1,1) model, as this model is representative, simple, and powerful (Hansen and Lunde 2005). GARCH (1,1) considers only one lag of the squared return and one lag of the conditional variance. The equation for the GARCH (1,1) model is as follows:

$$\Sigma_t^2 = \alpha_0 + \alpha_1 \varepsilon_{t-1}^2 + \beta_1 \sigma_{t-1}^2 \tag{9}$$

We used rolling forecast techniques, and the model used the training data as well as the past test data.

3.5. LSTM

The equations for LSTM are as follows:

The cell state (C):

$$C_t = f_t \times C_{t-1} + i_t \times \widetilde{C}_t \tag{10}$$

The forget gate (f):

$$f_t = \sigma\Big(W_f \cdot [h_{t-1}, x_t] + b_f\Big) \tag{11}$$

The input gate (i):

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \tag{12}$$

The output gate (o):

$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) \tag{13}$$

The hidden state h_t:

$$h_t = o_t \times tanh(C_t) \tag{14}$$

The candidate for cell state at timestamp t ($\widetilde{C}_t$):

$$\widetilde{C}_t = tanh(W_C \cdot [h_{t-1}, x_t] + b_C) \tag{15}$$

where

- σ: sigmoid activation function.
- W_f, W_i, W_C, and W_o: weight matrices for the forget gate, input gate, candidate values, and output gate, respectively.
- b_f, b_i, b_C, and b_o: biases corresponding to each gate.
- h_{t-1}: last hidden state.
- x_t: current input.
- tanh is the hyperbolic tangent activation function.

Our model used a 22-day windowed dataset, and we performed a hyperparameter search using Keras Random Search with 20 trials to optimize our results. Our hyperparameter search results suggest a two-layered bidirectional LSTM model with 128 and 32 units. We used 200 epochs with a batch size of 32 to train the LSTM model. We applied a dropout rate of 0.2 to reduce overfitting. We used bidirectional LSTM because it retains forward and backward sequence information, which helps the model better understand the context—a crucial aspect in forecasting volatility. We used the Adam optimizer to optimize our model, employed an early stopping with a patience level of 20, and only kept the best model. A visual for the LSTM architecture is shown in Figure 3. The details for the LSTM model are shown in Appendix B.

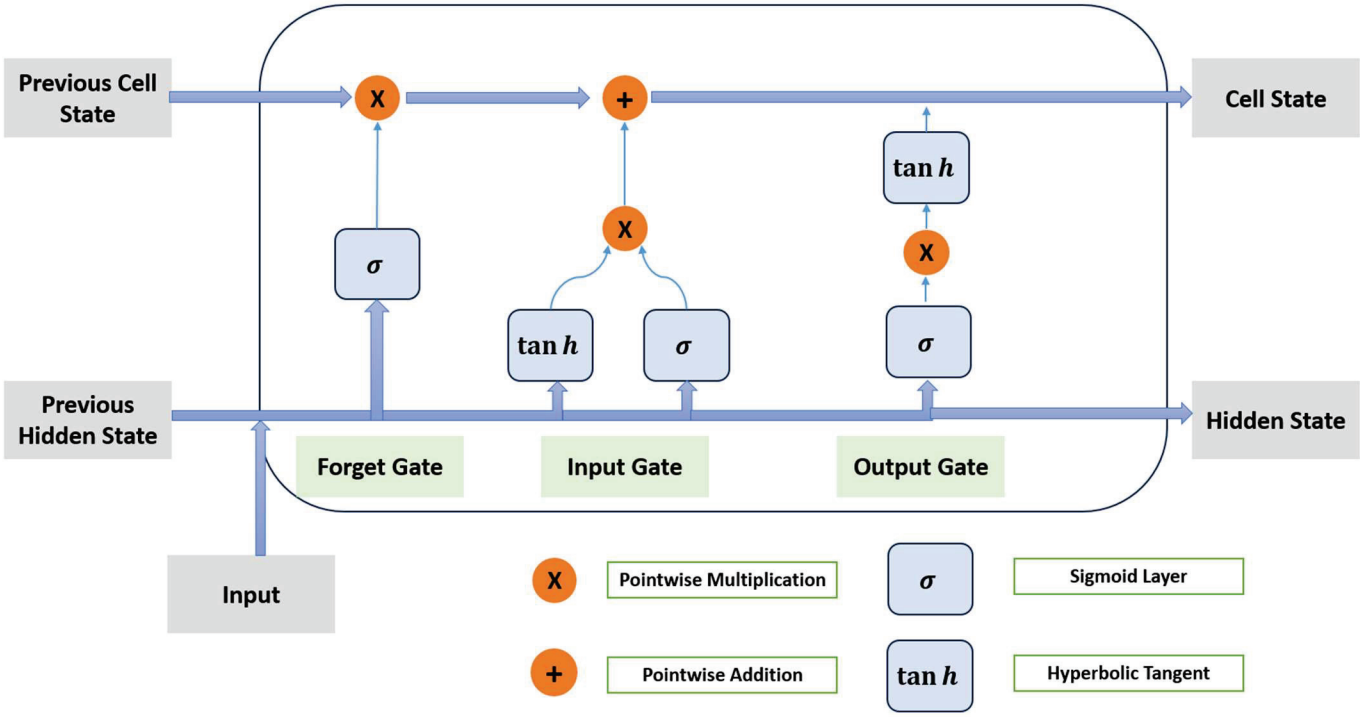


Figure 3. Long Short Term Memory.

3.6. Transformer

We built a vanilla transformer that is modified to work with time series tasks. We processed our data in batches of 32 over 200 epochs. During preprocessing, we windowed our dataset into 22-day segments and reshaped it into three dimensions. The details for the transformer model are shown in Appendix C.

We performed a hyperparameter search and set the hidden layer size to be 16, dropout to be 0.1, and the number of units in the multilayer perceptron (MLP) layer to be 256. We used 4 layers for our encoder. Each layer contains Layer Normalization, Multi-Head Attention, dropout, Residual Connection, and a feed-forward network. Figure 4 shows a visual for Multi-Head Attention. The equation for Multi-Head Attention is as follows:

$$\text{MultiHead}(Q, K, V) = \text{Concat}(\text{head}_1, \text{head}_2, \ldots, \text{head}_h)W_O \tag{16}$$

$$\text{head}_i = \text{Attention}\left(QW_{Q_i}, KW_{K_i}, VW_{V_i}\right) \tag{17}$$

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{18}$$

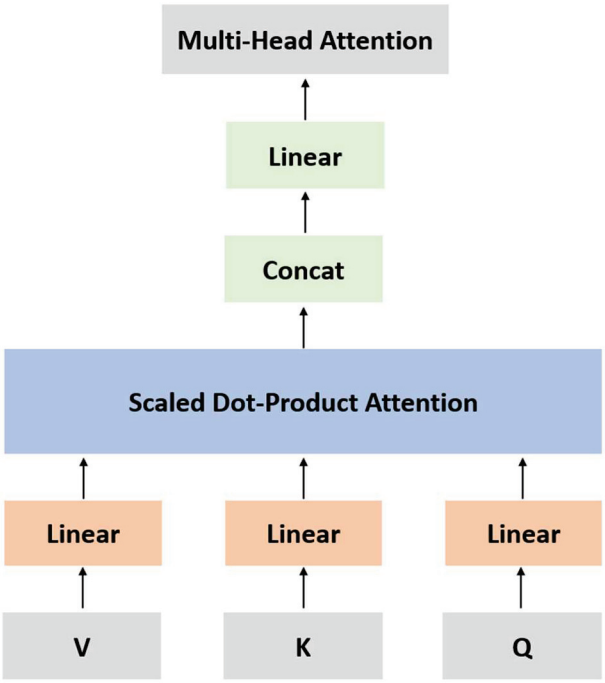


Figure 4. Multi-Head Attention (transformer).

3.7. Evaluation Metrics

To compute for errors, we used mean absolute error (MAE), root mean square error (RMSE), and root mean square percentage error (RMSPE):

$$\text{MAE} = \frac{1}{N}\sum_{i=1}^{N}|A_i - F_i| \tag{19}$$

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(A_i - F_i)^2} \tag{20}$$

$$\text{RMSPE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(\frac{A_i - F_i}{A_i}\right)^2} \tag{21}$$

where

- N: Total number of observations.
- A_i: Actual value for the ith observation.
- F_i: Predicted value for the ith observation.

4. Performance During Crisis Periods

4.1. Overview

We selected two recent crisis periods to test the models' performance during extreme scenarios where volatility spikes: the 2008 Financial Crisis and the 2020 COVID-19 Pandemic. We kept the other parameters in the models unchanged except for using different periods of training and testing data for each case. As demonstrated in the table below, LSTM showed slightly better performance than the transformer overall, while both outperformed the GARCH and IV models. All numerical results are shown in Table 2, and the baseline visual is shown in Figure 5.

Table 2. Comparison of different models' performance (the best performing model is **bolded**).

Model Name	RMSPE (Base)	RMSE (Base)	MAE (BASE)	RMSPE (2008)	RMSE (2008)	MAE (2008)	RMSPE (COVID10)	RMSE (COVID10)	MAE (COVID10)	RMSPE (COVID15)	RMSE (COVID15)	MAE (COVID15)
GARCH (1,1)	0.198	0.029	0.022	0.152	0.039	0.024	0.273	0.061	0.037	0.273	0.061	0.037
Implied Volatility	0.511	0.068	0.054	0.555	0.127	0.089	1.001	0.170	0.112	1.001	0.171	0.112
2-layered LSTM	**0.056**	**0.010**	**0.006**	**0.080**	**0.024**	**0.014**	**0.092**	**0.013**	**0.012**	**0.007**	**0.016**	**0.009**
Transformer	0.057	0.011	0.007	0.248	0.166	0.069	0.104	0.040	0.016	0.010	0.020	0.013

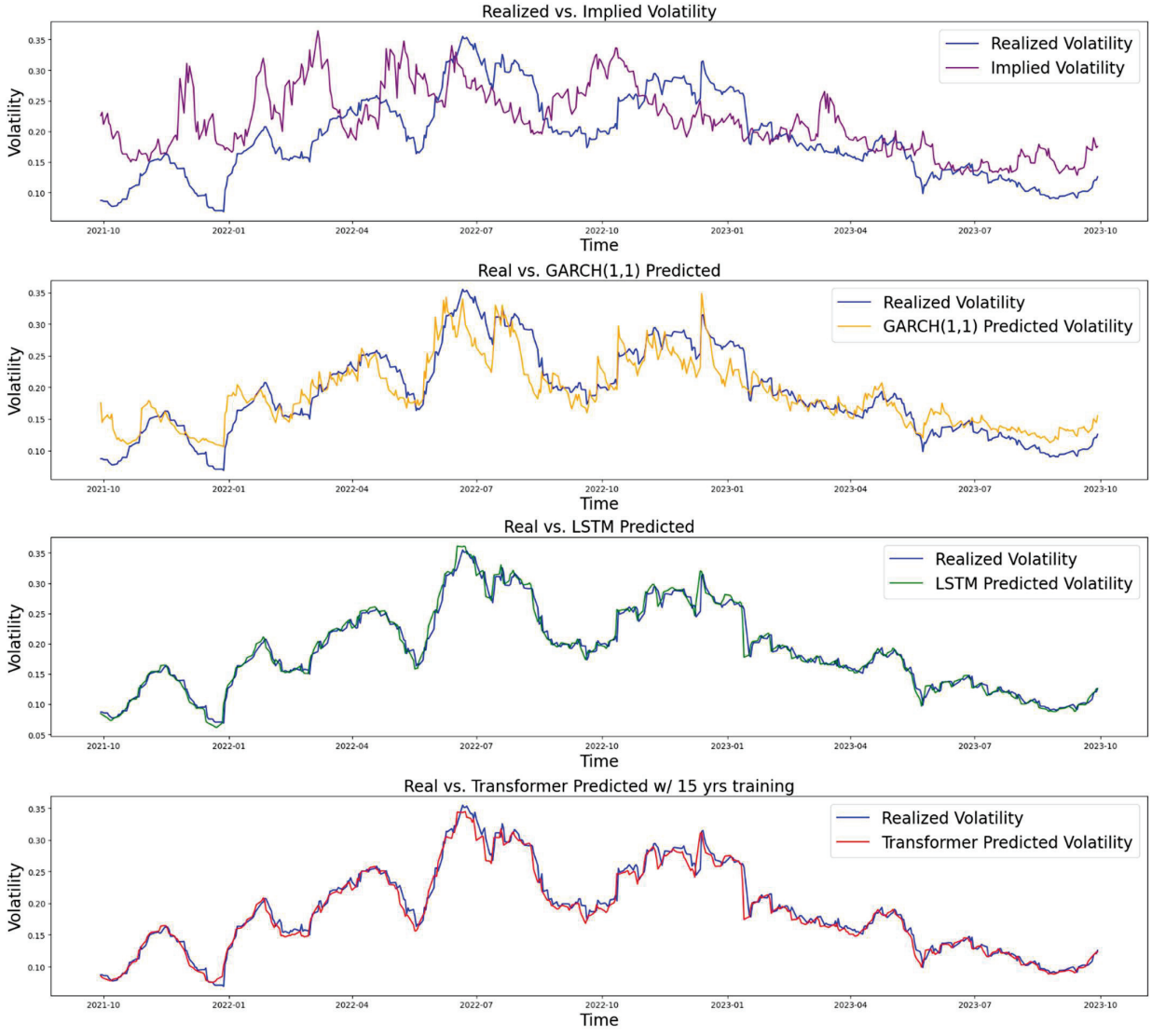


Figure 5. Comparison of performance (base scenario: training from 3 December 1993 to 28 September 2021).

4.2. 2008 Financial Crisis

We used a training period from 3 February 1997 to 8 February 2007 and a testing period from 9 February 2007 to 9 February 2009. The results are shown in Figure 6.

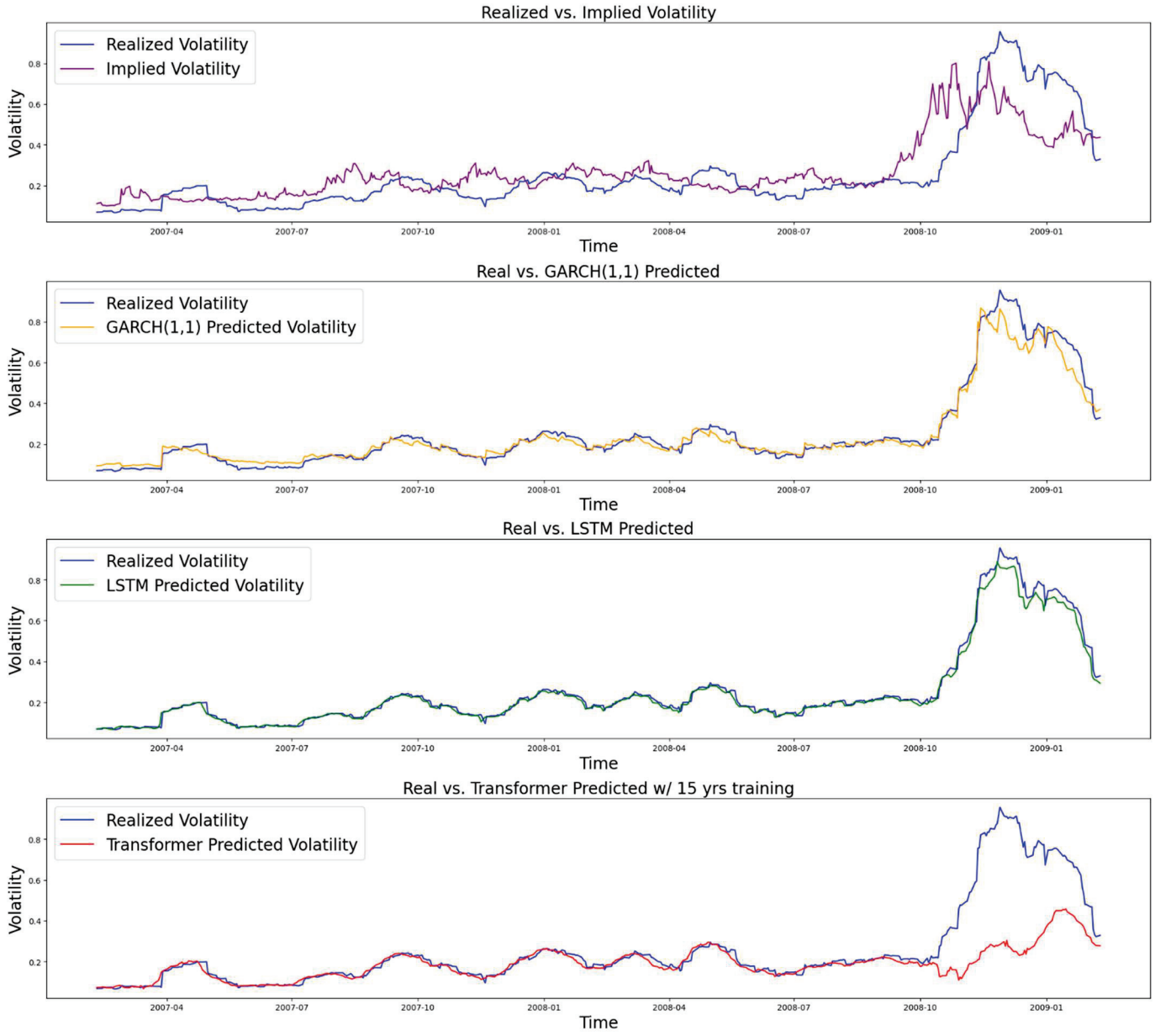


Figure 6. Comparison of performance (Financial Crisis period: training from 3 February 1997 to 8 February 2007).

4.3. 2020 COVID-19 Pandemic

We used two different data periods for the training part of the COVID-19 case to compare the performance of machine learning models when given different lengths of training data. Specifically, we selected two periods: one that does not include the 2008 Financial Crisis, and another that does. The first period uses data from 1 March 2010 to 3 March 2020 for training and from 4 March 2020 to 2 March 2022 for testing. The second period uses data from 25 February 2005 to 3 March 2020 for training and from 4 March 2020 to 2 March 2022 for testing. We refer to the two datasets as COVID10 and COVID15, respectively. The results are shown in Figures 7 and 8, respectively.

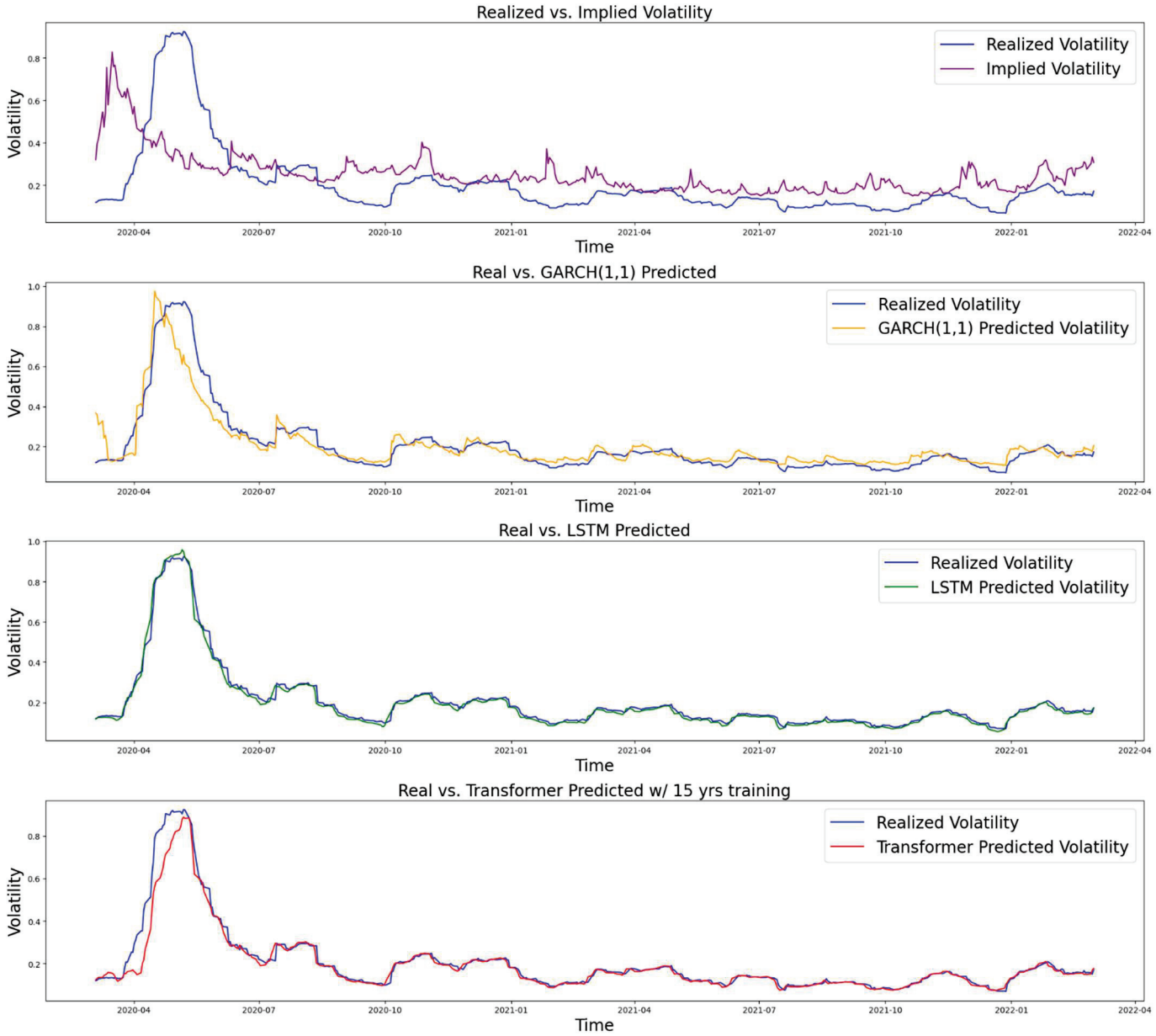


Figure 7. Comparison of performance (COVID with 10 years of training: from 1 March 2010 to 3 March 2020).

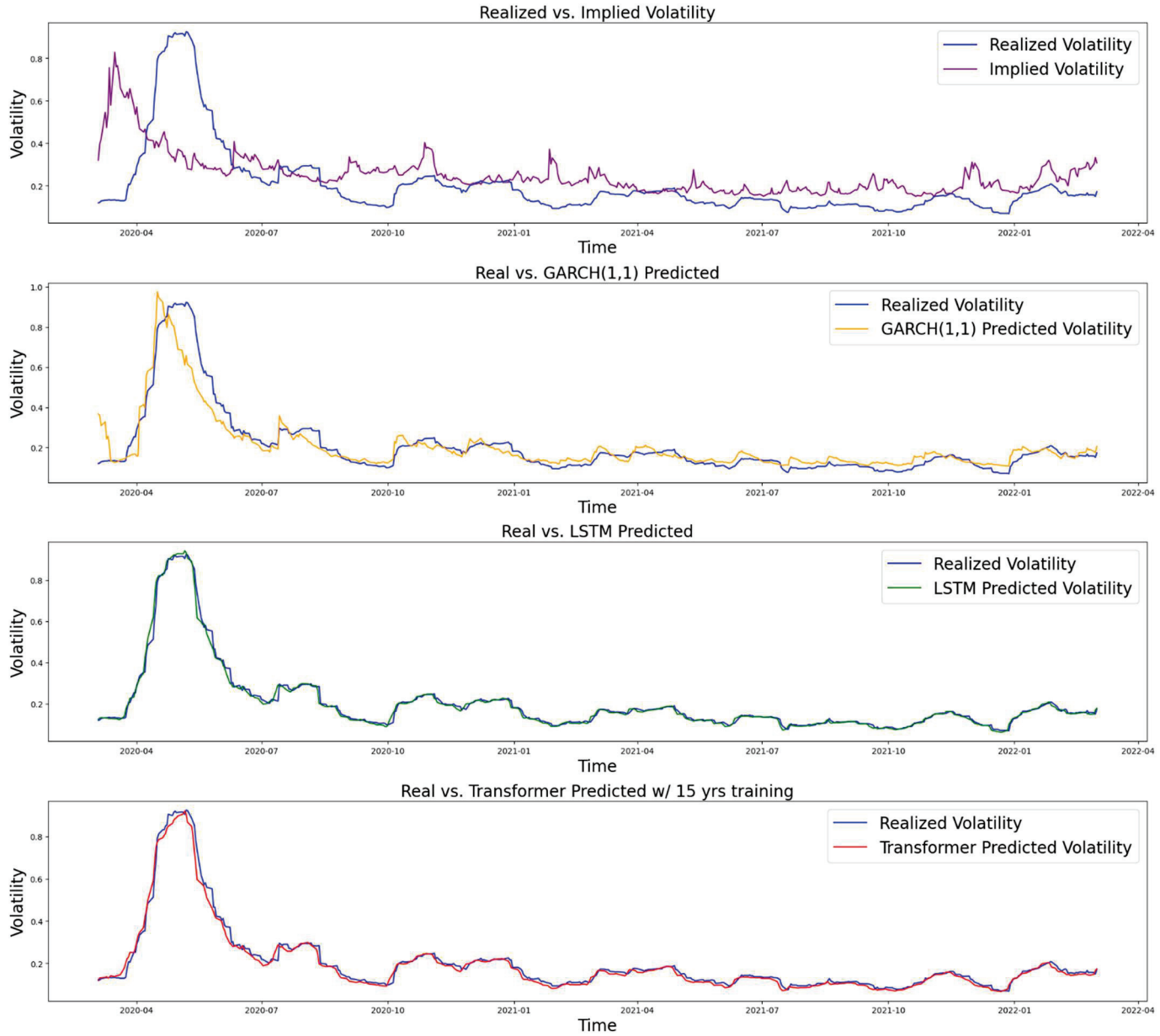


Figure 8. Comparison of performance (COVID with 15 years of training: from 25 February 2005 to 3 March 2020).

5. Results

As our goal is to forecast future market volatility, the value we are trying to predict is the realized volatility in the future. To compare the performance of our models, we calculate both the RMSE and the RMSPE between the model prediction and the target. We used historical VIX close value, which makes the computation faster than what would happen intraday in real-time, which is updated every 15 s.

In the thirty years of S&P 500 data we used to train and test our models, the LSTM and transformer models significantly outperformed the econometrics models, such as GARCH. Among these two models, the older LSTM performed slightly better with an RMSE of 0.010, while the transformer had an RMSE of 0.011. The results are shown in Table 2 and Figure 5. In Section 4, we conducted comparative studies to show how models perform when fewer data are given and when extreme events occur, like the 2008 Financial Crisis and the COVID-19 pandemic. While the LSTM model still outperforms, the transformer model's performance drops significantly, mainly due to the requirement for a large amount

of training data. This is most evident when testing during the financial crisis period of 2008 (Figure 6). The relatively short period of training data did not include a comparable period where stock prices were as volatile as they were in the financial crisis, and the transformer model failed to capture the sudden increase in volatility. As the LSTM model performs the best, we also tested similar models, such as GRU and RNN, which are simplified versions of the LSTM. They both performed reasonably well with an RMSE of 0.018 and 0.017, respectively, while there is still a significant gap compared with LSTM. We include the details for these two models in Appendix D.

To ensure universality and fairness in comparison, we constructed models for each category that are general and representative instead of fine-tuned for specific tasks. While these models provide an adequate representation for the purpose of this benchmark, further optimization could reveal the full potential of each model's performance.

6. Discussion

Volatility is an important measure for both regulators and financial institutions. Volatility is dynamically changing, and a time series analysis would help them better understand the risk they face at each time step. A good estimation of volatility provides a more accurate result in stress tests. Volatility is also an essential input for option pricing. A more accurate estimate will help traders to price options better.

Throughout the paper, we discussed how time series analysis has been applied to forecast volatility. The inherent nature of time series forecasting is to forecast the future using past information. However, given the vast amount of past information, we do not know which information to use and how to weigh its importance. From a historical perspective, statistical models made various assumptions, such as the clustering of volatility in GARCH models or the exponential weighting of recent information in the Exponential Weighted Moving Average. However, neural networks made a paradigm shift with minimal assumptions made. The process is now dependent on the data through backpropagation, as opposed to some explicit assumptions and formulas. Transformers revolutionized the process once again by abandoning the recurrence structure entirely and using the attention mechanism to determine the importance of past information in predicting the future.

As shown in our results section, the LSTM model performed the best in our datasets. However, the performance of different models also depends on the data on which they are trained. For example, a 30-year dataset is relatively small for the transformer model, which was originally used for large language processing tasks. This is part of why the transformer does not perform better than older models like LSTM. In addition, there were numerous changes in the stock market throughout the 30-year period; this evolving nature of the stock market made forecasting especially complex. Although the non-machine learning models perform worse than the ML models, they do not require a large dataset and are faster to train, which makes them advantageous when data are limited.

The data shortage problem discussed above can be mitigated through data augmentation or by using higher frequency data. For example, data augmentation can be performed through seasonal trend decomposition (Wen et al. 2019), applying transformations in the feature space (DeVries and Taylor 2017), and Generative Adversarial Networks (Esteban et al. 2017). Intraday market microstructure should be considered if using higher frequency data, such as minute-level data, because volatility tends to be higher in the first 30 and last 15 min of the trading day compared to other periods (Sampath and ArunKumar 2013). When implementing these models in practice, it is worth noting that machine learning models take significant time to train and make predictions. The transformer model takes approximately two hours and fifteen minutes to run on our machine using Nvidia's A100 GPU through Google Colab, while LSTM takes approximately one and a half hours. In-

corporating the latest available data into the model also takes additional time. Real-time trading faces the tradeoff of requiring extra time to include the new data. Furthermore, overfitting, missing data, and the quality of data are all challenges in using NNs to forecast volatility (Bhuiyan et al. 2025). Therefore, NNs need modifications before being applied to high-frequency trading, such as L1 and L2 regularization, using simpler models, or improved feature engineering (Karanam et al. 2018). Even though classical models like GARCH do not predict as accurately, their relatively simple calculation makes them more reliable in a real trading environment, while more complex machine learning models may suffer from slippage and generate results that deviate from their predictions in a real environment. Machine learning models are also challenging to interpret because their training does not follow an explicit rule (Rudin 2019). However, there is an ongoing trend in making machine learning more interpretable (Carvalho et al. 2019). For example, heatmaps and activation visualization can be used to visualize key characteristics influencing the model's prediction accuracy, and sensitivity analysis can show the sensitivity of each feature (Karanam et al. 2018).

7. Conclusions

Volatility is an essential part of financial institutions' risk exposure, which makes volatility forecasting an important task. A benchmark comparing the efficacy and performance of different forecasting techniques can be beneficial. While existing literature reviews focus on specific models, like GARCH, there remains a gap for a holistic and up-to-date assessment. Our study summarizes the key attributes of market volatility, such as its dynamic nature, clustering behavior, long memory, heavy tails, and the asymmetric relationship between prices and volatility. The study also offers a comprehensive review of volatility forecasting methods, ranging from traditional models to the current state of the art. Traditional models like GARCH have performed well; however, machine learning algorithms such as LSTMs and transformers have enhanced forecasting accuracy even further. Specifically, in the thirty-year dataset we use, the two-layered LSTM model produces an RMSPE of only 0.056, the transformer model produces an RMSPE of 0.057, while the GARCH model produces an RMSPE of 0.198, and the implied volatility produces even higher RMSPE. Similar results have been shown in testing with other time intervals, as detailed in the results section. This outperformance is a trend that's expected to advance as even more sophisticated algorithms emerge.

However, machine learning models also have their limitations. The amount of daily financial data is relatively small when compared to other machine learning applications. The lack of sufficient data may lead to undertraining of the models and their failure to predict sudden moves in volatility, as evident in the increase in RMSE in the models when training with a shorter period of data compared to the original 28 years. Furthermore, machine learning models demand significant computational resources and training times. Machine learning models have also been claimed as black boxes with results that are difficult to explain.

8. Future Works

Despite the inherent challenges in predicting volatility due to its sensitivity to a multitude of factors, including economic, corporate, psychological, and unforeseeable exogenous shocks, our research has shown that forecasting volatility is possible and that its accuracy is expected to increase as we employ more sophisticated models. It has been demonstrated that a combination of existing models, like GARCH and LSTM, can improve forecasting accuracy. We also believe the state-of-the-art models, such as Retentive Networks, hold potential for future applications in volatility prediction. Furthermore,

metrics outside of those traditionally applied to finance, such as distance-based metrics like Dynamic Time Wrapping, may be used to measure the accuracy of models.

On the data side, data augmentation or higher frequency data, such as hourly or minute data, may yield different results, especially for machine learning models whose performance is highly dependent on the amount of data. Using different indices or individual stocks to test their robustness and perform sensitivity analyses on features may also be beneficial. Furthermore, our models can be extended to forecast future prices to attract a larger audience. It can also be adjusted to forecast the volatility of other asset classes, such as commodities, interest rates, and cryptocurrencies.

Finally, machine learning models have long been criticized for being black boxes and hard to interpret. There are several ways to address this problem, as pointed out by Carvalho et al. (2019). There can be a summary of features, model internals, data points, or an approximation of the black box machine learning models using interpretable models.

Author Contributions: Conceptualization, Z.Q., C.K., F.S., and E.S.C.; Methodology, Z.Q., C.K., and E.S.C.; Software, Z.Q.; Validation, Z.Q.; Writing and Visualization, Z.Q., C.K., and E.S.C.; Supervision, C.K., F.S., and E.S.C.; Funding Acquisition, Z.Q. and E.S.C. All authors have read and agreed to the published version of the manuscript.

Funding: We would like to thank the Keck Foundation for their grant to Pepperdine University to support our research in Data Science.

Data Availability Statement: The data and code implementation are available at https://github.com/WithAnOrchid0513/VolData (accessed on 1 April 2025).

Acknowledgments: This work was supported in part by the Keck Institute Undergraduate Research Grant at Pepperdine University.

Conflicts of Interest: The authors declare no conflicts of interest.

Appendix A. Key Observations on Stylized Facts of Volatility

First, volatility is dynamic and displays temporal clustering (Mandelbrot 1997). Significant volatility today would suggest a higher likelihood of significant movement in the upcoming days (Kim and Shin 2023). Moreover, volatility's past fluctuations can exert lasting influences on its future path, signifying that volatility possesses a long memory (Poon and Granger 2003). Another observation is that the probability of extreme market events exceeds what the normal distribution would predict, indicating that return distributions exhibit heavy tails (Cont 2001).

Additionally, the leverage effect or the Asymmetric Volatility Phenomenon (AVP) suggests a negative correlation exists between prices and volatility: as prices drop, volatility intensifies, and as prices rise, volatility diminishes, though to a lesser extent (McAleer and Medeiros 2008; Aıt-Sahalia 2017; Engle and Ng 1993). Due to the AVP, option prices exhibit a skew. Options with strike prices below the market typically have higher implied volatility than their higher strike counterparts, which can be explained by several factors. Firstly, loss aversion suggests that investors tend to prioritize avoiding losses over achieving equivalent gains (Tversky and Kahneman 1991). Secondly, when a stock's value decreases, its financial leverage rises as the percentage of debt in its capital structure increases, making the stock riskier and boosting its volatility (Christie 1982). Lastly, adverse events increase conditional covariances substantially, whereas positive shocks have a mixed impact on conditional covariances (Bekaert and Wu 2000).

Volatility also exhibits mean reversion. Unlike stocks that have a positive drift, implied volatility tends to gradually increase before earnings and major events such as the Federal Open Market Committee (FOMC) meetings. It can also spike when encountering

unexpected events. However, in either case, volatility tends to revert to the mean after the event happens (Goudarzi 2013).

Appendix B. Notes on the LSTM Model

LSTM specifically utilizes sigmoid and tanh activation functions. The sigmoid function confines any input value within a range from 0 to 1, whereas the tanh function limits it between -1 and 1. With the current and previous information, these activation functions determine the amount of previous information to keep or discard in the forget gate. If the forget gate outputs 0, it forgets everything; if it outputs 1, it remembers everything. Then, these functions are used for the input and output gates. In the output gate, the short-term memory for the next period will be calculated using short-term memory. This will become the output for the current LSTM cell and the input for the next period. The long-term memory receives updates by initially processing the forgotten state and then assimilating it with the input state. This iterative updating of short-term and long-term memory persists till the model concludes its operation.

Appendix C. Notes on the Transformer Model

The transformer process begins with Layer Normalization, which normalizes the input data to have zero mean and unit variance. Then, the Multi-Head Attention calculates a weighted sum of the input based on its relationships with other parts of the input. This part can be computed in parallel to leverage GPU. Dropout is then applied to regularize the network. It achieves this by randomly setting a fraction of the input units to 0 at each update during training, which helps prevent overfitting.

The Residual Connection assists in counteracting the vanishing gradient problem encountered in deep networks. Finally, the feed-forward network uses 1D convolutional filters with a RELU activation function (Nair and Hinton 2005) that replaces the feed-forward layer in the original transformer. This adds nonlinearity and allows the model to learn complex patterns. We perform our hyperparameter search on the number of attention heads, their dimension, the hidden layer size in the feed-forward network, the number of encoder blocks, mlp units, and the dropout rate. We searched using Keras Random Search with 20 epochs.

Appendix D. GRU and RNN Models

We implemented a GRU model and an RNN model to show how models similar to LSTM perform. RNN is a starting point for this type of model, and we use it as a baseline to show how LSTM and GRU have improved based on it. As discussed in the Literature Review, RNN suffers from the vanishing gradient problem, which limits its ability to retain long-term dependencies. LSTM and GRU address this problem, and among them, GRU is a simplified version of LSTM with fewer parameters and faster training. We implemented a GRU model with a hyperparameter search on the number of layers, units, and dropout rate using Keras Random Search with 20 epochs. We also implemented a simple RNN as the baseline. In the same 30-year dataset, we obtained an RMSE of 0.018 for GRU and 0.017 for RNN, which are both worse than the LSTM model.

References

Abdel-Hamid, Ossama, Abdel-rahman Mohamed, Hui Jiang, Li Deng, Gerald Penn, and Dong Yu. 2014. Convolutional Neural Networks for Speech Recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 22: 1533–45. [CrossRef]

Ahmed, Sabeen, Ian E. Nielsen, Aakash Tripathi, Shamoon Siddiqui, Ghulam Rasool, and Ravi P. Ramachandran. 2022. Transformers in Time-Series Analysis: A Tutorial. *arXiv* arXiv:2205.01138. [CrossRef]

Andersen, Torben G., ed. 2018. *Volatility*. The International Library of Critical Writings in Economics 344. Northampton: Edward Elgar Publishing, Inc.

Andersen, Torben, Tim Bollerslev, Peter Christoffersen, and Francis Diebold. 2005. *Volatility Forecasting*. w11188. Cambridge: National Bureau of Economic Research. [CrossRef]

Asgharian, Hossein, Ai Jun Hou, and Farrukh Javed. 2013. The Importance of the Macroeconomic Variables in Forecasting Stock Return Variance: A GARCH-MIDAS Approach. *Journal of Forecasting* 32: 600–12. [CrossRef]

Aıt-Sahalia, Yacine. 2017. Estimation of the Continuous and Discontinuous Leverage Effects. *Journal of the American Statistical Association* 112: 1744–58. [CrossRef]

Bekaert, Geert, and Guojun Wu. 2000. Asymmetric Volatility and Risk in Equity Markets. *Review of Financial Studies* 13: 1–42. [CrossRef]

Bhuiyan, Md Shahriar Mahmud, Md Al Rafi, Gourab Nicholas Rodrigues, Md Nazmul Hossain Mir, Adit Ishraq, M. F. Mridha, and Jungpil Shin. 2025. Deep Learning for Algorithmic Trading: A Systematic Review of Predictive Models and Optimization Strategies. *Array* 26: 100390. [CrossRef]

Bildirici, Melike, and Özgür Ersin. 2015. Forecasting Volatility in Oil Prices with a Class of Nonlinear Volatility Models: Smooth Transition RBF and MLP Neural Networks Augmented GARCH Approach. *Petroleum Science* 12: 534–52. [CrossRef]

Black, Fischer, and Myron Scholes. 1973. The Pricing of Options and Corporate Liabilities. *The Journal of Political Economy* 81: 637–54. [CrossRef]

Bollerslev, Tim. 1986. Generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics* 31: 307–27. [CrossRef]

Bollerslev, Tim. 2008. Glossary to ARCH (GARCH). Available online: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=1263250 (accessed on 1 April 2025).

Bollerslev, Tim, and Michael Melvin. 1994. Bid—Ask Spreads and Volatility in the Foreign Exchange Market. *Journal of International Economics* 36: 355–72. [CrossRef]

Box, G. E. P., and David A. Pierce. 1970. Distribution of Residual Autocorrelations in Autoregressive-Integrated Moving Average Time Series Models. *Journal of the American Statistical Association* 65: 1509–26. [CrossRef]

Breiman, Leo. 2001. Random Forest. *Machine Learning* 45: 5–32. [CrossRef]

Carvalho, Diogo V., Eduardo M. Pereira, and Jaime S. Cardoso. 2019. Machine Learning Interpretability: A Survey on Methods and Metrics. *Electronics* 8: 832. [CrossRef]

CBOE. 2019. Volatility Index Methodology: Cboe Volatility Index. CBOE. Available online: https://cdn.cboe.com/api/global/us_indices/governance/Volatility_Index_Methodology_Cboe_Volatility_Index.pdf (accessed on 3 June 2024).

Chen, Tianqi, and Carlos Guestrin. 2016. XGBoost: A Scalable Tree Boosting System. Paper presented at 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13–17; pp. 785–94. [CrossRef]

Chow, T. W. S., and C. T. Leung. 1996. Neural Network Based Short-Term Load Forecasting Using Weather Compensation. *IEEE Transactions on Power Systems* 11: 1736–42. [CrossRef]

Christie, A. 1982. The Stochastic Behavior of Common Stock Variances Value, Leverage and Interest Rate Effects. *Journal of Financial Economics* 10: 407–32. [CrossRef]

Chung, Junyoung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. 2014. Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling. *arXiv* arXiv:1412.3555. [CrossRef]

Cirstea, Razvan-Gabriel, Chenjuan Guo, Bin Yang, Tung Kieu, Xuanyi Dong, and Shirui Pan. 2022. Triformer: Triangular, Variable-Specific Attentions for Long Sequence Multivariate Time Series Forecasting. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, 1994–2001*. Vienna: International Joint Conferences on Artificial Intelligence Organization. [CrossRef]

Connor, J. T., R. D. Martin, and L. E. Atlas. 1994. Recurrent Neural Networks and Robust Time Series Prediction. *IEEE Transactions on Neural Networks* 5: 240–54. [CrossRef]

Cont, Rama. 2001. Empirical Properties of Asset Returns: Stylized Facts and Statistical Issues. *Quantitative Finance* 1: 223. [CrossRef]

Derman, E. 1999. More Than You Ever Wanted to Know About Volatility Swaps. Available online: https://emanuelderman.com/wp-content/uploads/1999/02/gs-volatility_swaps.pdf (accessed on 1 May 2025).

DeVries, Terrance, and Graham W. Taylor. 2017. Dataset Augmentation in Feature Space. *arXiv* arXiv:1702.05538. [CrossRef]

Diamond, Richard V. 2012. VIX as a Variance Swap. *SSRN Electronic Journal*. [CrossRef]

Engle, Robert. 2004. Risk and Volatility: Econometric Models and Financial Practice. *American Economic Review* 94: 405–20. [CrossRef]

Engle, Robert F. 1982. Autoregressive Conditional Heteroscedasticity with Estimates of the Variance of United Kingdom Inflation. *Econometrica* 50: 987. [CrossRef]

Engle, Robert F., and Victor K. Ng. 1993. Measuring and Testing the Impact of News on Volatility. *The Journal of Finance* 48: 1749–78. [CrossRef]

Esteban, Cristóbal, Stephanie L. Hyland, and Gunnar Rätsch. 2017. Real-Valued (Medical) Time Series Generation with Recurrent Conditional GANs. *arXiv* arXiv:1706.02633. [CrossRef]

Fang, Tong, Tae-Hwy Lee, and Zhi Su. 2020. Predicting the Long-Term Stock Market Volatility: A GARCH-MIDAS Model with Variable Selection. *Journal of Empirical Finance* 58: 36–49. [CrossRef]

García-Medina, Andrés, and Ester Aguayo-Moreno. 2024. LSTM–GARCH Hybrid Model for the Prediction of Volatility in Cryptocurrency Portfolios. *Computational Economics* 63: 1511–42. [CrossRef]

Ge, Wenbo, Pooia Lalbakhsh, Leigh Isai, Artem Lenskiy, and Hanna Suominen. 2023. Neural Network–Based Financial Volatility Forecasting: A Systematic Review. *ACM Computing Surveys* 55: 1–30. [CrossRef]

Goudarzi, Hojatallah. 2013. Volatility mean reversion and stock market efficiency. *Asian Economic and Financial Review* 3: 1681.

Granger, C. W. J., and Roselyne Joyeux. 1980. An introduction to long-memory time series models and fractional differencing. *Journal of Time Series Analysis* 1: 15–29. [CrossRef]

Gu, Wentao, Suhao Zheng, Ru Wang, and Cui Dong. 2020. Forecasting Realized Volatility Based on Sentiment Index and GRU Model. *Journal of Advanced Computational Intelligence and Intelligent Informatics* 24: 299–306. [CrossRef]

Gudelek, M. Ugur, S. Arda Boluk, and A. Murat Ozbayoglu. 2017. A Deep Learning Based Stock Trading Model with 2-D CNN Trend Detection. In *2017 IEEE Symposium Series on Computational Intelligence (SSCI)*. Honolulu: IEEE, pp. 1–8. [CrossRef]

Hansen, Peter R., and Asger Lunde. 2005. A Forecast Comparison of Volatility Models: Does Anything Beat a GARCH(1,1)? *Journal of Applied Econometrics* 20: 873–89. [CrossRef]

He, Kaijian, Qian Yang, Lei Ji, Jingcheng Pan, and Yingchao Zou. 2023. Financial Time Series Forecasting with the Deep Learning Ensemble Model. *Mathematics* 11: 1054. [CrossRef]

Hochreiter, Sepp, and Jürgen Schmidhuber. 1997. Long Short-Term Memory. *Neural Computation* 9: 1735–80. [CrossRef] [PubMed]

Holt, Charles C. 2004. Forecasting Seasonals and Trends by Exponentially Weighted Moving Averages. *International Journal of Forecasting* 20: 5–10. [CrossRef]

Hornik, Kurt, Maxwell Stinchcombe, and Halbert White. 1989. Multilayer Feedforward Networks Are Universal Approximators. *Neural Networks* 2: 359–66. [CrossRef]

Huang, Cheng-Zhi Anna, Ashish Vaswani, Jakob Uszkoreit, Noam Shazeer, Ian Simon, Curtis Hawthorne, Andrew M. Dai, Matthew D. Hoffman, Monica Dinculescu, and Douglas Eck. 2018. Music Transformer. *arXiv* arXiv:1809.04281.

Huang, Dashan, Fuwei Jiang, Kunpeng Li, Guoshi Tong, and Guofu Zhou. 2022. Scaled PCA: A New Approach to Dimension Reduction. *Management Science* 68: 1678–95. [CrossRef]

Hull, John. 2018. *Options, Futures, and Other Derivatives*, 10th ed. New York: Pearson.

Hurst, H. E. 1951. Long-Term Storage Capacity of Reservoirs. *Transactions of the American Society of Civil Engineers* 116: 770–99. [CrossRef]

Johnston, F R, J E Boyland, M Meadows, and E Shale. 1999. Some Properties of a Simple Moving Average When Applied to Forecasting a Time Series. *Journal of the Operational Research Society* 50: 1267–71. [CrossRef]

Karanam, Raghunath Kashyap, Vineel Mouli Natakam, Narasimha Rao Boinapalli, Narayana Reddy Bommu Sridharlakshmi, Abhishekar Reddy Allam, Pavan Kumar, SSMLG Gudimetla Naga Venkata, Hari Priya Kommineni, and Aditya Manikyala. 2018. Neural Networks in Algorithmic Trading for Financial Markets. *Asian Accounting and Auditing Advancement* 9: 115–26.

Kelly, Bryan T., and Dacheng Xiu. 2023. Financial Machine Learning. *SSRN Electronic Journal* 13: 205–363. [CrossRef]

Kim, Donggyu, and Minseok Shin. 2023. Volatility Models for Stylized Facts of High-frequency Financial Data. *Journal of Time Series Analysis* 44: 262–79. [CrossRef]

Kim, Ha Young, and Chang Hyun Won. 2018. Forecasting the Volatility of Stock Price Index: A Hybrid Model Integrating LSTM with Multiple GARCH-Type Models. *Expert Systems with Applications* 103: 25–37. [CrossRef]

Kitaev, Nikita, Łukasz Kaiser, and Anselm Levskaya. 2020. Reformer: The efficient transformer. *arXiv* arXiv:2001.04451. [CrossRef]

Knight, Frank. 1921. *Risk, Uncertainty and Profit*. Boston: Houghton Mifflin Company.

Kolen, John F., and Stefan C. Kremer. 2009. Gradient Flow in Recurrent Nets: The Difficulty of Learning LongTerm Dependencies. In *A Field Guide to Dynamical Recurrent Networks*. Piscataway: IEEE. [CrossRef]

Koo, Eunho, and Geonwoo Kim. 2022. A Hybrid Prediction Model Integrating GARCH Models With a Distribution Manipulation Strategy Based on LSTM Networks for Stock Market Volatility. *IEEE Access* 10: 34743–54. [CrossRef]

Li, Shiyang, Xiaoyong Jin, Yao Xuan, Xiyou Zhou, Wenhu Chen, Yu-Xiang Wang, and Xifeng Yan. 2019. Enhancing the Locality and Breaking the Memory Bottleneck of Transformer on Time Series Forecasting. In *Advances in Neural Information Processing Systems 32* (NeurIPS 2019). Available online: https://dl.acm.org/doi/10.5555/3454287.3454758 (accessed on 1 May 2025).

Liu, Ze, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. 2021. Swin Transformer: Hierarchical Vision Transformer Using Shifted Windows. *arXiv* arXiv:2103.14030.

Loh, Wei-Yin. 2011. Classification and Regression Trees. *WIREs Data Mining and Knowledge Discovery* 1: 14–23. [CrossRef]

Ludvigson, Sydney C., and Serena Ng. 2007. The Empirical Risk–Return Relation: A Factor Analysis Approach. *Journal of Financial Economics* 83: 171–222. [CrossRef]

Mandelbrot, Benoit B. 1997. The Variation of Certain Speculative Prices. In *Fractals and Scaling in Finance*. Edited by Benoit B. Mandelbrot. New York: Springer, pp. 371–418. [CrossRef]

Marcek, Dusan. 2018. Forecasting of Financial Data: A Novel Fuzzy Logic Neural Network Based on Error-Correction Concept and Statistics. *Complex & Intelligent Systems* 4: 95–104. [CrossRef]

McAleer, Michael, and Marcelo C. Medeiros. 2008. Realized Volatility: A Review. *Econometric Reviews* 27: 10–45. [CrossRef]

Nair, Vinod, and Geoffrey E Hinton. 2005. Rectified Linear Units Improve Restricted Boltzmann Machines. Available online: https://dl.acm.org/doi/10.5555/3104322.3104425 (accessed on 1 May 2025).

Nie, Yuqi, Nam H. Nguyen, Phanwadee Sinthong, and Jayant Kalagnanam. 2023. A Time Series Is Worth 64 Words: Long-Term Forecasting with Transformers. *arXiv* arXiv:2211.14730.

Olorunnimbe, Kenniy, and Herna Viktor. 2024. Ensemble of Temporal Transformers for Financial Time Series. *Journal of Intelligent Information Systems* 62: 1087–111. [CrossRef]

Ozdemir, Ali Can, Kurtuluş Buluş, and Kasım Zor. 2022. Medium- to Long-Term Nickel Price Forecasting Using LSTM and GRU Networks. *Resources Policy* 78: 102906. [CrossRef]

Pomerleau, Dean A. 1988. ALVINN, an Autonomous Land Vehicle in a Neural Network. In *Advances in Neural Information Processing Systems 1* (NIPS 1988). Available online: https://papers.nips.cc/paper/1988/hash/812b4ba287f5ee0bc9d43bbf5bbe87fb-Abstract.html (accessed on 1 May 2025).

Poon, Ser-Huang, and Clive Granger. 2003. Forecasting Volatility in Financial Markets: A Review. *Journal of Economic Literature* 41: 478–539. [CrossRef]

Pranav, B. M., and Vinay Hegde. 2021. Volatility Forecasting Techniques Using Neural Networks: A Review. *International Journal of Engineering Research & Technology (IJERT)* 10: 748–52.

Qian, Bo, and Khaled Rasheed. 2004. Hurst exponent and financial market predictability. In *IASTED Conference on Financial Engineering and Applications*. Cambridge: IASTED International Conference.

Raudys, Aistis, and Edvinas Goldstein. 2022. Forecasting Detrended Volatility Risk and Financial Price Series Using LSTM Neural Networks and XGBoost Regressor. *Journal of Risk and Financial Management* 15: 602. [CrossRef]

Rudin, Cynthia. 2019. Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead. *Nature Machine Intelligence* 1: 206–15. [CrossRef]

Sampath, Aravind, and G. ArunKumar. 2013. Do Intraday Volatility Patterns Follow a 'U' Curve? Evidence from the Indian Market. *SSRN Electronic Journal*. [CrossRef]

Sezer, Omer Berat, and Ahmet Murat Ozbayoglu. 2018. Algorithmic Financial Trading with Deep Convolutional Neural Networks: Time Series to Image Conversion Approach. *Applied Soft Computing* 70: 525–38. [CrossRef]

Sezer, Omer Berat, Mehmet Ugur Gudelek, and Ahmet Murat Ozbayoglu. 2020. Financial Time Series Forecasting with Deep Learning: A Systematic Literature Review: 2005–2019. *Applied Soft Computing* 90: 106181. [CrossRef]

Shiller, Robert J. 1999. *Market Volatility*. 1. paperback ed. [Nachdr.]. Cambridge: MIT Press.

Stock, James H, and Mark W Watson. 2002. Forecasting Using Principal Components From a Large Number of Predictors. *Journal of the American Statistical Association* 97: 1167–79. [CrossRef]

Sun, Yutao, Li Dong, Shaohan Huang, Shuming Ma, Yuqing Xia, Jilong Xue, Jianyong Wang, and Furu Wei. 2023. Retentive Network: A Successor to Transformer for Large Language Models. *arXiv* arXiv:2307.08621.

Taylor, James W. 2004. Smooth Transition Exponential Smoothing. *Journal of Forecasting* 23: 385–404. [CrossRef]

Tversky, A., and D. Kahneman. 1991. Loss Aversion in Riskless Choice: A Reference-Dependent Model. *The Quarterly Journal of Economics* 106: 1039–61. [CrossRef]

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2023. Attention Is All You Need. *arXiv* arXiv:1706.03762.

Venugopal, V., and W. Baets. 1994. Neural Networks and Statistical Techniques in Marketing Research: A Conceptual Comparison. *Marketing Intelligence & Planning* 12: 30–38. [CrossRef]

Wen, Qingsong, Jingkun Gao, Xiaomin Song, Liang Sun, Huan Xu, and Shenghuo Zhu. 2019. RobustSTL: A Robust Seasonal-Trend Decomposition Algorithm for Long Time Series. *Proceedings of the AAAI Conference on Artificial Intelligence* 33: 5409–16. [CrossRef]

Wen, Qingsong, Tian Zhou, Chaoli Zhang, Weiqi Chen, Ziqing Ma, Junchi Yan, and Liang Sun. 2023. Transformers in Time Series: A Survey. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*. Macau: International Joint Conferences on Artificial Intelligence Organization, pp. 6778–86. [CrossRef]

Whaley, Robert E. 2009. Understanding the VIX. *The Journal of Portfolio Management* 35: 98–105. [CrossRef]

Wu, Haixu, Jiehui Xu, Jianmin Wang, and Mingsheng Long. 2021. Autoformer: Decomposition Transformers with Auto-Correlation for Long-Term Series Forecasting. *Advances in Neural Information Processing Systems* 34: 22419–30.

Zeng, Ailing, Muxi Chen, Lei Zhang, and Qiang Xu. 2023. Are Transformers Effective for Time Series Forecasting? *Proceedings of the AAAI Conference on Artificial Intelligence* 37: 11121–28. [CrossRef]

Zerveas, George, Srideepika Jayaraman, Dhaval Patel, Anuradha Bhamidipaty, and Carsten Eickhoff. 2021. A Transformer-Based Framework for Multivariate Time Series Representation Learning. Paper presented at 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining, Singapore, August 14–18; pp. 2114–24. [CrossRef]

Zhang, Chao, Yihuang Zhang, Mihai Cucuringu, and Zhongmin Qian. 2023. Volatility Forecasting with Machine Learning and Intraday Commonality. *arXiv* arXiv:2202.08962.

Zhang, Yunhao, and Junchi Yan. 2023. Crossformer: Transformer utilizing cross-dimension dependency for multivariate time series forecasting. Paper presented at Eleventh International Conference on Learning Representations, Kigali, Rwanda, May 1–5.

Zhao, Pengfei, Haoren Zhu, Wilfred Siu Hung NG, and Dik Lun Lee. 2024. From GARCH to Neural Network for Volatility Forecast. *arXiv* arXiv:2402.06642. [CrossRef]

Zheng, Yi, Qi Liu, Enhong Chen, Yong Ge, and J. Leon Zhao. 2014. Time Series Classification Using Multi-Channels Deep Convolutional Neural Networks. In *Web-Age Information Management*. Edited by Feifei Li, Guoliang Li, Seung-won Hwang, Bin Yao and Zhenjie Zhang Science. Lecture Notes in Computer. Cham: Springer International Publishing, vol. 8485, pp. 298–310. [CrossRef]

Zhou, Haoyi, Shanghang Zhang, Jieqi Peng, Shuai Zhang, Jianxin Li, Hui Xiong, and Wancai Zhang. 2021. Informer: Beyond Efficient Transformer for Long Sequence Time-Series Forecasting. *Proceedings of the AAAI Conference on Artificial Intelligence* 35: 11106–15. [CrossRef]

risks

MDPI

Article

Credit Risk Prediction Using Machine Learning and Deep Learning: A Study on Credit Card Customers

Victor Chang [1,*], Sharuga Sivakulasingam [1], Hai Wang [2], Siu Tung Wong [3], Meghana Ashok Ganatra [1] and Jiabin Luo [1]

[1] Department of Operations and Information Management, Aston Business School, Aston University, Birmingham B4 7ET, UK; sharuga1225@gmail.com (S.S.); meghana.ganatra@gmail.com (M.A.G.); j.luo2@aston.ac.uk (J.L.)
[2] School of Computer Science and Digital Technologies, Aston University, Birmingham B4 7ET, UK; h.wang10@aston.ac.uk
[3] Institute of Finance and Technology, University College London, London WC1E 6BT, UK; tommywong962@gmail.com
* Correspondence: v.chang1@aston.ac.uk or victorchang.research@gmail.com

Abstract: The increasing population and emerging business opportunities have led to a rise in consumer spending. Consequently, global credit card companies, including banks and financial institutions, face the challenge of managing the associated credit risks. It is crucial for these institutions to accurately classify credit card customers as "good" or "bad" to minimize capital loss. This research investigates the approaches for predicting the default status of credit card customer via the application of various machine-learning models, including neural networks, logistic regression, AdaBoost, XGBoost, and LightGBM. Performance metrics such as accuracy, precision, recall, F1 score, ROC, and MCC for all these models are employed to compare the efficiency of the algorithms. The results indicate that XGBoost outperforms other models, achieving an accuracy of 99.4%. The outcomes from this study suggest that effective credit risk analysis would aid in informed lending decisions, and the application of machine-learning and deep-learning algorithms has significantly improved predictive accuracy in this domain.

Keywords: credit risk prediction; credit risk; classification; machine learning

1. Introduction

Credit cards offer an easy method of borrowing money to pay for a range of goods and services in the modern era. Credit cards function as a replacement for cash and debit cards and are also widely used in daily shopping. Moreover, credit cards have become indispensable to the contemporary economy for many individuals. From the view of financial institutions and banks, they must decide whether to approve credit card applications when a customer submits the application. Different factors are considered by creditors when determining whether the customer is good or bad in terms of risk and repayment. According to practical scenarios, our approach categorizes a "bad customer" as one whose credit has been past due for 60 days or more, whereas a "good customer" is one whose debt has been past due for less than 60 days.

Credit scores are a general risk governance strategy in banks and other credit institutions, which utilize the personal data of credit card customers to signify the likelihood of future default. They help banks or financial institutions to determine whether it is appropriate to offer credit cards to applicants. Evading bad or default customers is essential for banks and other financial companies to avoid unwanted costs.

Banks and other credit card centers use a credit scoring system to determine how risky it is to lend to applicants. The application form filled out by the customer is rich in information and a valuable asset for the institution, helping to assign the credit score and

finalize the card issuance. A cutoff point will be established by creditors for credit scoring. The institution might opt not to lend to the applicant if the score falls below the cutoff; otherwise, it would charge more from the applicant upon issuing a card.

Credit risk assessment is crucial to commercial banks' credit risk management. The danger of a borrower defaulting on their loan due to the inability to make regular payments is called credit risk. It shows the chance that credit card loan providers might not receive the interest or principal owed to them on time. Commercial banks can efficiently avoid credit risk and make better lending decisions by accurately assessing the credit risk of their applicants. The key to favorable portfolio quality is an efficient underwriting and loan approval procedure, and one of the primary responsibilities of the function is to minimize unnecessary risks.

> A lucrative credit card client portfolio demands effective credit card customer segmentation. It serves as a risk management tool and enables the company to provide clients with appropriate offers and gain a better understanding of their credit card portfolios. Credit card users can be classified into five categories based on how they use their cards: max or full payers, revolvers, non-payers, traders, and non-users. According to Sajumo (Sajumon 2015), among these five categories, revolvers are the key clients for businesses, as credit card corporations can make profits at their expense. They only make the minimum payment required, if any, and continue using their cards as usual to make purchases. They are not influenced by high-interest rates. Any credit card provider would prefer to avoid having non-payers as customers. They obtain all accessible credit cards, utilize all available credit on those cards, but fail to make any payments.

As mentioned, from a company's perspective, it is crucial to understand the customers' backgrounds and financial conditions before providing them with a credit facility to maintain a healthy credit portfolio. This understanding will lead to an increase in the cardholder's lifespan and help to maintain a sustainable income. However, failure to effectively manage credit risk can lead to losses for credit card issuers and significantly impact their cash flow. Such disruptions can prevent credit card issuers from effectively reinvesting their resources, hindering growth and stability. Moreover, attempting to collect payments from non-payers presents its own set of challenges. Significant time and resources must be dedicated to this task, leading to additional operational expenses. These costs can be substantial, considering the expenses associated with employing collection agencies and other resources. Our research stands at the intersection of this critical challenge. By employing advanced machine-learning and deep-learning algorithms, we aim to more accurately classify customers into different categories based on their likelihood to default. This data-driven, automated approach allows banks or credit card issuers to manage their credit risk more effectively and efficiently, minimizing losses and ensuring a healthier cash flow.

Secondly, the financial landscape is not static. It is continuously affected by various events, including COVID-19, inflation, and recessions, which can destabilize customers' financial stability and consequently increase the rate of NPAs (Non-Performing Assets). Double-digit NPAs are a risk, and most banks and financial institutions struggle to maintain single-digit NPAs. In today's volatile economic climate, unmanaged or poorly managed credit risk exposure can lead to severe financial distress, even bankruptcy, if left unchecked. The domino effect of inaccurate credit risk management can be catastrophic, impacting not only the business itself but also stakeholders and the broader financial ecosystem. Therefore, in such an environment, it is essential not only to evaluate the credit risk of cardholders, but also to monitor them regularly. However, continuous monitoring can lead to the duplication of work and overburdening analysts, resulting in reduced efficiency for financial institutions. The importance of our research is further heightened. It streamlines the entire process, enabling financial institutions to effortlessly pre-process their credit customer data, select and apply credit risk classifiers, and accurately and efficiently predict the potential credit risk of the customers, and classifying them into "good" and "bad"

categories. This allows credit card issuers to adapt rapidly to changing circumstances and maintain their NPA rates within a manageable range.

Thirdly, maintaining good customers will increase trust in credit card products and boost product demand. A robust customer base allows institutions to upsell and cross-sell other banking products, fostering business growth. From a financial perspective, maintaining good customers leads to a healthier P&L (profit and loss account). Therefore, in the long-term approach, this research will help credit card issuers to achieve sustainable strategic growth.

The main goal of this article is to identify the riskiest category, or non-payers, using machine-learning and deep-learning techniques. Non-payers pose the most significant risk to the fund. Despite of the risks, credit card issuers will not stop issuing cards. Instead, card issuers will implement proactive risk management strategies and use more sophisticated and efficient credit management methods.

This paper discusses predicting the possibility of the default status of the credit customers using different methodologies such as random forest, neural networks, AdaBoost, eXtreme gradient boosting (XGBoost), light gradient boosting (LightGBM), and logistic regression to determine the best-performing algorithm based on accuracy, precision, recall, F1 score, ROC curve (receiver operating characteristic) and AUC score (area under the curve), and Matthews correlation coefficient (MCC). This approach will help financial institutions to make the correct decisions in identifying the right customers for credit card issuing.

1.1. Aims and Objectives

The main goal of the research is to create an automated method that can predict the consumer's default status based on each customer's application.

The following is a list of the objectives that we aim to achieve through the analysis:

- Determine the most essential characteristics that can be used to anticipate the defaulting status.
- Implement balancing techniques to enhance the appropriateness of the data for identifying and examining credit card data.
- Investigate various machine-learning and deep-learning methods and use credit card data as a predictive base.
- Identify and analyze the performance metrics that are the most appropriate for measuring the classification problem.
- Assess the robustness of the selected models over time to ensure sustained accuracy and reliability in varying economic conditions.

1.2. Research Contributions

Our research outputs have three key contributions as follows:

- Performing data analysis and visualization to gain insights, as well as summarizing significant data features to achieve a better understanding of the dataset. The approved and disbursed loan data from a given time period were studied, and the performance window was then selected and used to predict the performance of future periods (Han et al. 2012).
- Among six machine-learning algorithms, XGBoost was identified as the best-performing model and was chosen as the final model. Based on credit card customer information, a model was developed to classify good and bad customers. Refer to Section 5.3: Summary of Performance Metrics (Chen and Guestrin 2016).
- To understand the relationship between the dependent variables, a correlation matrix was employed, and feature importance was used to identify the critical features that determine the classifier's performance. Age, income, employment duration, and the number of family members are the primary predictors for the best-performing XGBoost model. Refer to Section 5.2: Feature Importance of the Best Performing Model XGBoost (Lundberg and Lee 2017).

- Evaluating the robustness of the models over time to ensure that accuracy remains consistent across different datasets and economic conditions. This ensures that the models are not only accurate on the initial dataset but also maintain performance as new data become available (Krawczyk 2016).

The paper is structured as follows: The first part introduces the research (Section 1), explaining the study's background and aim. The next part reviews previous relevant research (Section 2), followed by the data and methodology (Section 3). Section 4 discusses the different machine-learning algorithms used in this study. Section 5 presents the implementation and detailed results. Finally, Sections 6 and 7 provide the discussions and conclusion.

2. Related Literature

2.1. Importance of Credit Risk Analysis

Credit risk analysis plays pivotal roles in the business, finance, retail, and insurance industries. New techniques and technologies are crucial to help develop a better process of analysis and identify any potential issues.

Granting loans to applicants is a crucial concern for commercial banks worldwide (Xia et al. 2018). These financial organizations carefully assess the creditworthiness of their customers to prevent significant loss in the event of default. The fierce market rivalry compels them to separate the "good" applicants from the "bad" ones. Consequently, credit scoring has emerged as a popular research topic among scholars and financial institutions due to its effectiveness as a tool for assessing credit risk. According to (Sariannidis et al. 2020), financial institutions and credit analysts could benefit even more from developing machine learning-based techniques such as SVC and random forest in the future. Using quantitative and operational data to more precisely pinpoint the credit risk categories that their clients fall under would allow for a more accurate assessment of the customer's creditworthiness. To better understand and monitor the loan portfolios of banks and to pursue credit policies effectively, it is helpful to categorize the characteristics of clients and assign them to distinct credit risk categories.

The work of (Bao et al. 2019) discussed that recent studies have concentrated on the ensemble strategy, which incorporates various ML models for credit scoring. One of the more often used approaches is constructing consensus classification decisions based on the results of individual ML models. The work of (Chen et al. 2021) suggested that big data could be explored and analyzed using data analytics, which could help banks to reduce risks and make better investment decisions with reliable returns. The work of (Chang et al. 2020) described that using several different financial models could yield more precise results based on various scenarios, including investment requirements. The stakeholders could then be presented with all the outputs, increasing the likelihood of risk mitigation or avoidance.

According to (Buchanan and Wright 2021), machine learning is a significant factor influencing the financial services sector to a greater extent. Additionally, they looked at how machine learning and artificial intelligence are used in the UK financial services industry. They examined the UK's present AI/ML environment and concluded that credit scoring, financial distress prediction, robo-advising, and algorithmic trading are a few domains where machine learning has had a significant influence. They also noted that applications of ML for predicting credit market defaults are becoming more popular. ML may be used to evaluate character and reputation characteristics when predicting future payment patterns.

2.2. Methodologies for Credit Risk Analysis

There has been an increasing number of studies that have studied different machine-learning methodologies and justified them as favorable methods to calculate credit risk. These methodologies and their advantages from various machine-learning techniques are highlighted below.

Some of these research studies have focused on the development of one single algorithm; for example, a multinomial logistic regression model is used in the research by (Adha et al. 2018) to learn about the variables influencing default and attrition occurrences on credit. The accuracy of the multinomial logistic regression model in identifying customers based on the chance of defaulting is 95.3%.

Most of the research studies have applied several algorithms to compare the efficiency of the models, under different scenarios of credit risk modelling. The study by (Ullah et al. 2018) discussed that most card users, regardless of their ability to pay, misuse their credit cards and accrue cash-card debt. The biggest problem facing cardholders and banks right now is this issue. The study aimed to employ knowledge discovery in data to forecast credit card applicants' probability of defaulting on payment. Six regression approaches were used to identify credit default payment and card users, including K-nearest neighbors, the logistic regression model, SVM regression, AdaBoost, and random forest. Compared to other data mining approaches, AdaBoost performs best, having an accuracy rate of 88%.

The work of (Dm and Mm 2018) described that financial organizations must forecast loan defaults to reduce losses from non-payment. Their outcomes demonstrated that the support vector machine model outperformed the logistic regression model (accuracy: 86.12%, precision: 0.7831). The study advised financial institutions to use support vector machines to anticipate loan defaults.

According to (Ma et al. 2018), since its debut in 2016, LightGBM has been extensively adopted in the field of big data and machine learning. Together with XGBoost, it is considered a high-powered machine-learning tool. Publicly available experimental data indicate that LightGBM is more efficient and precise than other existing boosters. LightGBM is more precise, needs less memory, and is quicker than XGBoost. Further, experiments indicate that the LightGBM algorithm can acquire linear acceleration by utilizing many machines for specific training. Consequently, the benefits of this algorithm manifest in the following five aspects: low memory usage, rapid training speed, good model precision, support for parallel learning, and rapid processing of large datasets.

LightGBM (LGBM)'s dependability and adaptability will substantially facilitate the creation of a credit rating system. The research project conducted by (Naik 2021) aims develop an up-to-date credit scoring model that is contemporary in predicting credit defaults for unguaranteed loans such as credit cards using machine-learning approaches. According to the research findings, the LGBM classifier model is superior to other models in terms of its capacity to provide faster learning speeds, improve efficiency, and manage larger amounts of data effectively. With the highest accuracy of 95.5% and an AUC of 0.99, the LGBM surpasses the other models, which include logistic regression, SVM, K-nearest neighbors (KNN), XGBoost, decision trees, and random forest.

As per the paper by (Zhu et al. 2019), the loan default prediction model using the random forest algorithm can adapt to build a model to predict loan default in the given data compared with the other three methodologies, i.e., decision tree, logistic regression, and support vector machine. According to their experiment, the random forest model outperformed the other models with an accuracy of 98%, an AUC of 0.983, an F1 score of 0.98, and a recall of 0.99. Moreover, they added that the random forest algorithm works rapidly on large databases and is the most accurate algorithm compared with the other three algorithms. It can also deal effectively with errors in unbalanced data during the classification problem. Finally, it is a valuable method for assessing missing data, which can produce good accuracy even when a significant portion is missing. According to (Sayjadah et al. 2018), by measuring the customer's level of risk and using the model results, banks and financial institutions can advise businesses on making smart decisions. Their article examines the efficacy of credit card default forecasting. Random forest, logistic regression, and decision trees are used to analyze variables in predicting credit default, with random forest demonstrating a superior accuracy and area under the curve. As per the results, the random forest, with the highest accuracy of 82% and a better AUC of 0.77, best captures the criteria.

In the work of (Tian et al. 2020), models were compared and discussed based on their accuracy, AUC, and F1 score using suitable data cleaning and important feature selection. The gradient boosting decision tree is one of the top models, with an outstanding accuracy of 92.19%, AUC value of 0.97, and F1 score of 91.83%, when compared to logistic regression, decision trees, SVM, neural networks, random forest, and AdaBoost. The work demonstrated that the mentioned model had the finest ability for generalization and classification.

The work of (Duan 2019) suggests an MLP consisting of three hidden layers used to train the technique of backpropagation for loan default prediction in lending. It is demonstrated that the MLP model's approach classifies test data with 93% accuracy, which is more significant than the prediction accuracy of 75% gained using the MLP model with one hidden layer, logistic regression, SVM, AdaBoost, and decision trees. In the research carried out by (Bindal and Chaurasia 2018), the five data mining techniques logistic regression, naive Bayes, decision trees, MLP classifier (neural networks), and KNN were compared. The MLP classifier gave the best performance with an area ratio of 0.88. Also, logistic regression helped to identify the main components necessary for analyzing a customer's credit risk.

According to (Liu 2022), the backpropagation neural network (BPNN) can learn on its own, adapt on its own, acquire knowledge, and cope with uncertainty successfully. His research demonstrates that the neural network efficiently regulates individual credit administration, lowers credit risks for banks and financial institutions, and offers a new framework for decision-making for the banks' customer credit operations. The work of (Sun and Vasarhelyi 2018) shows how deep learning can be used to forecast credit card delinquencies. Deep neural networks outperform typical artificial neural networks, decision trees, naive Bayes, and logistic regression in terms of overall predictive performance and have the greatest overall accuracy (99.5%), F scores (0.7064), AUC (0.9547), and precision (0.8502). Also, they added that deep learning has successfully been applied, suggesting that AI has a lot of promise to assist and predict credit risk assessment by modeling for banks and financial institutions.

The work of (Wang et al. 2022) compared and analyzed three classification algorithms: decision trees, K-nearest neighbors, and XGBoost. The individual's credit risk evaluation algorithm based on XGBoost performs better in terms of the Type II error rate (0.199), accuracy (87.1%), and AUC (0.943). They concluded that the XGBoost-based model for assessing the probability of default on a personal credit line has a strong default discrimination capability and robustness.

The article by (Lin et al. 2023) stated that credit scoring models might still be unable to identify consumers who are unable or unwilling to make loan payments, leading to early loan defaults that would result in significant losses for lenders. Based on real-world credit data obtained from online lending mediums, their study tries to classify those bad customers who default on their loans soon after they are issued. They carried out experimental research based on various conditions of early defaults. The outcomes show that standard logistic regression is significantly outperformed by LightGBM regarding prediction ability for the classification job. Although the benefit is negligible regarding 1/N recall rates, ML-LightGBM performs even better in terms of AUC than a Bayesian-optimized LightGBM model. They concluded that ML-LightGBM is a favorable method for credit scoring and fraud detection. The study by (Ma et al. 2018) classifies and analyzes the Lending Club's loan dataset to forecast whether the customer will fail to make the payment in the future using the LightGBM and XGBoost algorithms. Each model's output is utilized to summarize the results. The study concluded that LightGBM's classification prediction results for the identical dataset are superior to those of XGBoost.

In (Guégan and Hassani 2018), the authors introduce the concept of "Regulatory Learning" in their study, focusing on the supervision of machine-learning models in the context of credit scoring. The paper emphasizes the importance of building models that not only achieve high predictive accuracy but also comply with regulatory standards—an

essential requirement in the financial industry. By applying various machine-learning algorithms to credit scoring, the authors illustrate how interpretability, transparency, and robustness of models are crucial for regulatory compliance. The study provides a comprehensive framework for integrating machine-learning techniques into credit scoring while considering the dynamic nature of risk and the need for periodic model validation. This work lays the groundwork for developing credit risk models that balance performance with regulatory requirements, a critical aspect that aligns with the focus of our research.

In another related study (GeeksforGeeks n.d), the authors conducted an in-depth analysis of credit risk using a variety of machine-learning and deep-learning models, including neural networks, logistic regression, AdaBoost, XGBoost, and LightGBM. The authors compared the performance of these models based on evaluation metrics such as accuracy, precision, recall, F1 score, ROC, and MCC. Their findings revealed that XGBoost showed a strong predictive performance, which aligns with our study's results.

In summary, a comparison of the different methodologies from the literature review is shown in Table 1.

Table 1. Accuracy comparison from the literature review.

Study	Model	Accuracy
Zhu et al. (2019)	Random forest	98.00%
Sayjadah et al. (2018)	Random forest	82.00%
Tian et al. (2020)	Gradient boosting decision tree	92.19%
Sun and Vasarhelyi (2018)	Neural network	99.50%
Duan (2019)	Neural network backpropagation (3 hidden layers)	93.00%
Wang et al. (2022)	XGBoost	87.10%
Naik (2021)	LGBM	95.50%
Adha et al. (2018)	Logistic regression	95.30%
Ullah et al. (2018)	AdaBoost	88.00%
Dm and Mm (2018)	SVM	86.12%

2.3. Reinforcement Learning in Finance

Reinforcement learning in banking: In recent times, there has been a notable surge in the utilization of reinforcement learning (RL) in the banking industry. In portfolio management, Q-learning and policy gradient approaches have demonstrated promise by enabling dynamic asset allocation strategies that adjust to shifting market conditions. In both bull and bear markets, Lucarelli and Borrotti (Lucarelli and Borrotti 2020) showed how a deep Q-network can be more effective at managing a portfolio than traditional approaches. The work of (Sumiea et al. 2024) investigated policy gradient approaches, which provide a more straightforward means of optimizing portfolio performance indicators and demonstrate particular strength in managing transaction costs and market effects.

Another area in RL applications is adaptive credit scoring systems based on multi-armed bandit algorithms. These techniques, like those presented by (Ali et al. 2024), enable credit scoring models to be continuously learned from and adjusted in response to new data. This strategy is beneficial in dynamic lending markets where macroeconomic conditions and borrower behavior change quickly.

However, there are several difficulties when applying RL in dynamic financial situations. The work of (Malibari et al. 2023) draws attention to several problems, including the non-stationarity of financial time series, the requirement for a substantial quantity of training data, and the challenge of defining suitable reward functions in intricate financial systems. Furthermore, the exploration–exploitation trade-off in reinforcement learning can

present a significant challenge in financial applications where exploration may result in actual financial losses.

2.4. Optimization Techniques

Evolutionary algorithms have demonstrated significant potential in selecting features for credit risk models. The work of (Xu and Zhang 2024) illustrated the efficacy of genetic algorithms in selecting optimal feature subsets for credit scoring, thereby enhancing model performance and reducing dimensionality. These methods are especially advantageous in high-dimensional financial datasets, where conventional feature selection methods may be computationally unfeasible.

In the field of financial forecasting, gradient-based methods continue to be the foundation of neural network training. Recent developments, including those proposed by (Behera et al. 2023), incorporate second-order optimization techniques and adaptive learning methods that considerably enhance model performance and the speed of convergence in predicting financial time series.

Regulatory compliance in risk assessment has become more critical due to the increasing importance of constrained optimization approaches. The work of (Maldonado et al. 2017) developed a constrained optimization framework for credit scoring that directly integrates regulatory requirements into the model optimization process.

Based on the above discussions, it can be observed that machine-learning models are becoming very efficient tools in credit risk scoring, and each of the models has its own advantages over others. Among the most applied algorithms, we particularly choose neural networks, logistic regression, AdaBoost, XGBoost, and LightGBM for their ability of learning, adapting, and predicting as approved in previous studies, to apply in the scenario of credit card customers in our considered case study. The results can benefit the credit card industry by developing appropriate algorithms to predict credit risks, in order to analyze the uncertainty in credit scoring and thus support the decision-making process to mitigate the effects from potential risks.

3. Data Analysis

3.1. Dataset Description and Methodology Overview

The research is based on a secondary dataset from Kaggle. This dataset is actual bank data presented on the website after removing the customers' sensitive personal information. There are two different data files, application_record.csv and credit_record.csv. The first application_record dataset includes the applicants' data, which may be used as predictive characteristics. The second one, the credit_record dataset, keeps track of consumers' credit card usage habits (credit history). The ID is the connecting column (primary key) of the application and credit record datasets. Two tables that are linked by ID were combined to create the data. The application_record.csv contains the columns of the client ID, gender, ownership of a car, phone, work phone, mobile phone, email, and property, number of children, family size, annual income, income category, educational level, marital status, housing type, birthday, and occupation type. Client ID, record month, and customer status are the columns in the credit_record.csv dataset. As a starting point, the record month is the month from which the data were collected; counting backward, 0 represents the current month, -1 represents the prior month, and so on. The status column lists the following amounts as past due: 0: 1–29 days, 1: 30–59 days, 2: 60–89 days, 3: 90–119 days, 4: 120–149 days, 5: write-offs for greater than 150 days that are past-due or bad debts. "C" indicates that month's repayment, and "**X**" indicates no loan for the month. Detailed descriptions of the variables are given in Appendix A.

An overview of the methodology of this study is shown in Figure 1, which presents the 5 steps of the analysis that are adopted in the following sections. Firstly, the dataset was tested and cleaned to improve the quality of the work; then exploratory data analysis was conducted to learn more about the data before pre-processing. After that, all six methodologies were used to differentiate the predicted results from the actual results

using the confusion matrix. During the process, each methodology was evaluated and compared using performance metrics such as accuracy, recall, precision, F1 score, ROC–AUC, and MCC.

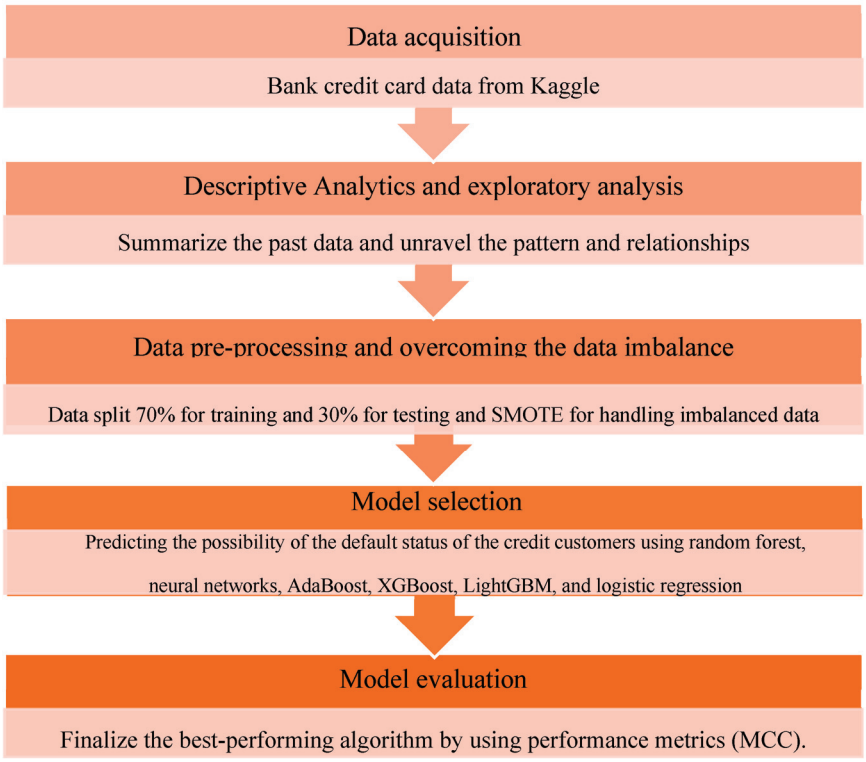


Figure 1. Methodology framework.

3.2. Exploratory Data Analysis (EDA)

Raw data, also known as unprocessed data, are only helpful if there is something to be gained from investigating it. EDA involves analyzing and visually representing the data to gain insights, as well as summarizing significant data properties to gain a better understanding of the dataset.

According to IBM (Education 2020) and (Aswini et al. 2020), EDA gives users a more profound knowledge of the variables in the data collection and their relationships. It is generally used to explore what data might reveal beyond the formal modeling or hypothesis testing assignment. EDA can also aid in determining if the statistical approaches being considered for the research methodology are appropriate.

Some models have a significant number of features, which can cause the arrangement and training processes to take more time and consume a significant amount of system memory. It requires considerable time and effort for each feature to scan through the various data instances and estimate every potential split point, which is the primary factor contributing to this behavior. It is evident that when there are extra features in the data, the efficiency and scalability of the model are far from optimal. It is recommended to have fewer characteristics to save time during the computing process and boost the model's performance.

The work of (Al-qerem et al. 2019) mentioned that data preparation is a crucial step when developing a classification model, as it significantly affects model accuracy. Applying feature selection techniques to a vast dataset is also of great importance; it improves accuracy and performance.

The summary statistics provided in Figure 2 help to understand the variable distribution more effectively.

Based on the observations from the above table, firstly, the variables were further analyzed, and "Flag mobile" was decided to be removed from the model prediction since

the minimum and maximum value are both "1". Secondly, outliers need to be checked for a few variables since there is a considerable difference between the maximum value and the 75th percentile value. Thirdly, the variables with the value "0" for a min, 25th, 50th, and 75th percentiles were further investigated.

	ID	CNT_CHILDREN	AMT_INCOME_TOTAL	DAYS_BIRTH	DAYS_EMPLOYED	FLAG_MOBIL	FLAG_WORK_PHONE	FLAG_PHONE	FLAG_EMAIL	CNT_FAM_MEMBERS
count	438557.000	438557.000	438557.000	438557.000	438557.000	438557.000	438557.000	438557.000	438557.000	438557.000
mean	6022176.270	0.427	187524.286	-15997.905	60563.675	1.000	0.206	0.288	0.108	2.194
std	571637.023	0.725	110086.853	4185.030	138767.800	0.000	0.405	0.453	0.311	0.897
min	5008804.000	0.000	26100.000	-25201.000	-17531.000	1.000	0.000	0.000	0.000	1.000
25%	5609375.000	0.000	121500.000	-19483.000	-3103.000	1.000	0.000	0.000	0.000	2.000
50%	6047745.000	0.000	160780.500	-15630.000	-1467.000	1.000	0.000	0.000	0.000	2.000
75%	6456971.000	1.000	225000.000	-12514.000	-371.000	1.000	0.000	1.000	0.000	3.000
max	7999952.000	19.000	6750000.000	-7489.000	365243.000	1.000	1.000	1.000	1.000	20.000

Figure 2. Summary statistics.

Unstructured data are transformed through data visualization into groupings and metrics that may be quickly used as smart business information for quick and effective decision-making. The application submitted date and the status value for each month following the open month for the credit card help to analyze the credit behavior. The credit card customers' past credit history can be compared during the various application months.

Additionally, it identifies relationships and patterns, as well as areas that work well or can be improved. The distribution of the variables and the data balance are examined using a variety of data visualization approaches such as bar charts, pie charts, and histograms to enhance the comprehension of the information and its aspects.

Unique values of each column were checked to identify the redundant columns. When the attribute has numerous unique values, it takes longer to conclude the data analysis, which should focus on the crucial part of our research questions. The detailed data analysis followed the steps below.

3.2.1. Performance Window and Target Variable Creation

Following the issuance of the customer's credit card, the details of the customer will be retained in the system, and each transaction will be monitored for various reasons. One of the most critical functions of monitoring the account's credit status is to track how much money is being spent and how much is being paid back. To maintain a credit scorecard for each month, the status must be categorized as either good or bad for each customer.

According to the algorithm, the bank or financial institution can distinguish between good and bad customers using the available features from the application filled out by the customer. In the context of this study, "bad" accounts are those whose overdue balances have a status of 2, 3, 4, or 5, while "good" accounts are all other types of accounts. The dataset needs to include information regarding the opening date of the credit card. When analyzing the data, the month with the earliest "MONTHS_BALANCE" is deemed to be the account's opening month. Then, we reorganize the data such that month 0 represents the beginning month, and one month after the beginning month is indicated by month 1, see Figure 3.

Figure 4 below depicts the monthly distribution of accounts by status. It shows each account-opening month, beginning with month 1 and continuing through month 60, and helps to review the performance of the portfolio. From the time window, it is necessary to identify their status and accounts according to the month they were created.

Over the course of all account-opening months, the bad rate ratio needs to be computed for the entire portfolio to locate the stable period of the bad rate. In the beginning, there was only a modest increase in the number of credit cards; however, this may have been insignificant for the models. Figure 4 shows that accounts that have been open for more than 50 months demonstrate a dramatic increase in the percentage of bad loans.

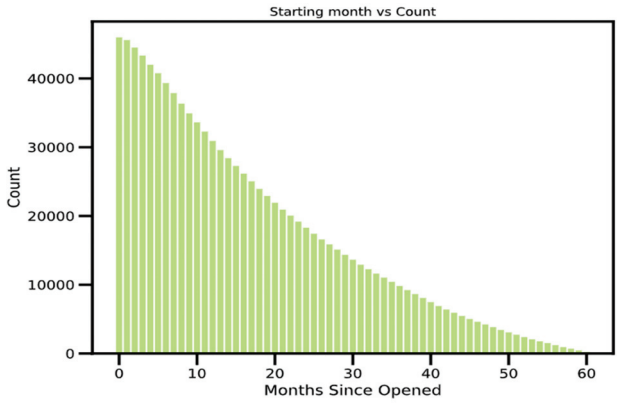


Figure 3. Performance over the period.

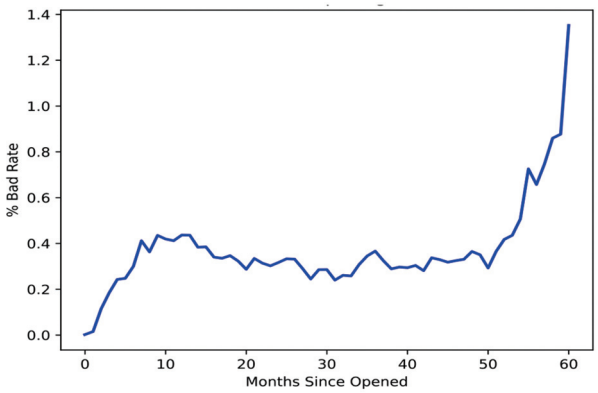


Figure 4. Bad rate of the portfolio.

As previously indicated, the status column contained various values that had been converted to binary numbers. The status values 2, 3, 4, and 5 are changed to 1, while the others are set to 0. A "bad customer" is one whose past-due balance exceeds 59 days, while a "good customer" is one whose past-due amount is less than 60 days. According to the binary value, 1 represents a bad customer, whereas 0 represents a good one.

As per the time series graph shown in Figure 5, the bad rate has nearly stabilized after one year (12 months). Based on this, the first 12 months can be considered as a performance window or the time frame for the analysis. Any customers that go delinquent within the first year will be labeled as "bad," while the remainder will be considered "good". Customers are categorized as bad or good based on their status throughout the initial 12 months, as shown in Figure 5.

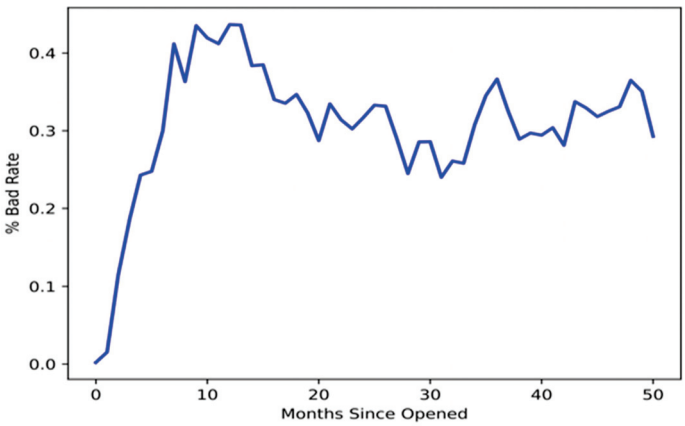


Figure 5. Bad rate of the portfolio after one year.

3.2.2. Target Distribution

According to Figure 6 below, the data are exceptionally imbalanced, with a rate of 1.3% for bad customers.

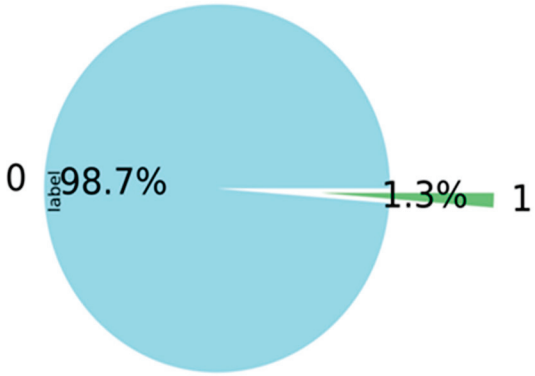


Figure 6. Target distribution.

3.2.3. Handling Outliers

Outliers can affect the quality of the research since all statistical data are susceptible to their effects, including means, standard deviations, and all other statistical inferences based on them. Handling outliers is one of the essential steps in data pre-processing. According to Figure 7 below, there are outliers that need to be removed from the data regarding days employed, total income, count of children, and count of family members.

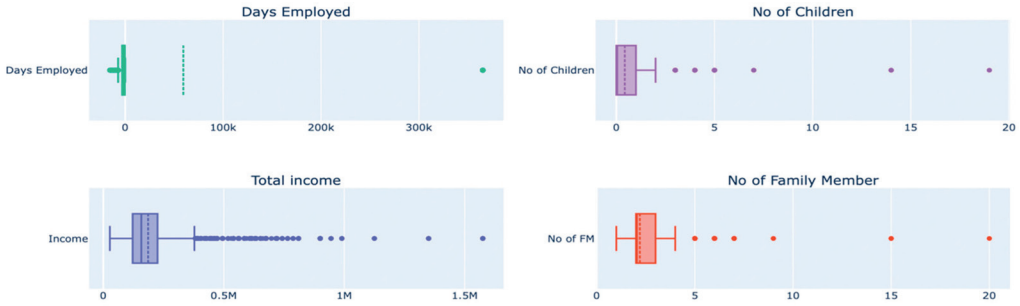


Figure 7. Box plots of features.

Outliers for the columns' days employed, the number of children, total income, and number of family members are detected using box plots, as shown above, and handled using the IQR method. The IQR method is used to detect the outliers by setting up a bar outside Q1 and Q3. Weights that fall beyond the fence range are concluded as outliers. To construct the fence results, we multiplied the IQR by 1.5, subtracted this amount from Q1, and added it to Q3. Outliers are any observations that exceed 1.5 IQR below Q1 or more than 1.5 IQR above Q3. Then, outliers were removed to ensure they did not impact the model's outcomes, as shown in Figure 8.

The missing values were analyzed, and 32% of the values were missing in the variable "OCCUPATION_TYPE". With the account-opening month and date of birth, the age and months of experience as of the application date are calculated backward from the application date to the date of birth and from the application date to the months of experience, respectively.

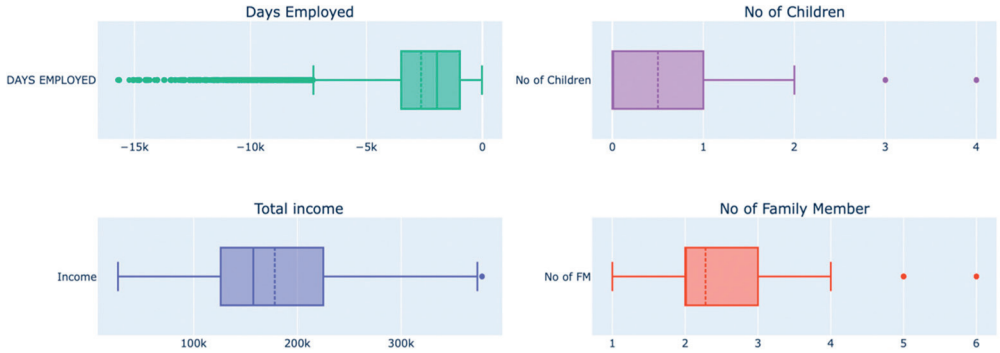


Figure 8. Box plots of features after removing outliers.

3.2.4. Data Visualization

The boxen plot and box plot, as displayed in Figure 9, are used to compare and understand the distribution of essential variables and the status of the customer.

Figure 9. Distribution of important variables.

The observations from the graphs above are as follows:

According to the first graph, the younger population, in general, poses a greater risk than the older population. It is evident that banks and other financial institutions should consider a person's age when deciding whether to issue them a credit card.

- The second graph shows that people with less experience typically pose a greater risk than those with more experience. When issuing a card, banks and other financial organizations should consider experience as one of the factors.
- As per the third graph, the red square refers to the average income, showing that the average income of the bad customers is below the average income of the good customers.
- It is evident from the pie charts shown in Figure 10 below that good customers have a higher proportion of property ownership compared to bad customers.

The histograms in Figure 11 show the distribution of the numeric variables for the most important variables:

- The income total shows the highest distribution between 50,000 and 275,000.
- Age shows the highest distribution between 300 months and 600 months, which is 25 to 50 years.
- Employment months have the highest distribution between 0 and 50 months, less than 1 year to 4 years.

The correlation matrix, illustrated in Figure 12 below, is a chart that displays the correlation coefficients established between different sets of variables. The table shows the correlation between each variable and each value. This makes it possible to identify the pairs with the highest correlation. In the study, a correlation matrix is initially generated to determine how each pair of variables relates to one another. It should be considered to remove variables if there is a significant relationship between them; hence, a set of highly correlated characteristics will not contribute any new information or very little. Moreover, they will complicate the algorithm and increase the possibility of errors. It is beneficial to remove the variables that are significantly correlated with one another in order to reduce

memory and speed issues. According to the confusion matrix presented below in Figure 12, the number of children and family members is highly correlated with a correlation of 85%. Initially, it was decided to remove one of the strongly correlated variables; however, both variables are kept for analysis because the correlation is less than 90%. Also, even after keeping both variables in the research, there is no adverse impact on the performance metrics results.

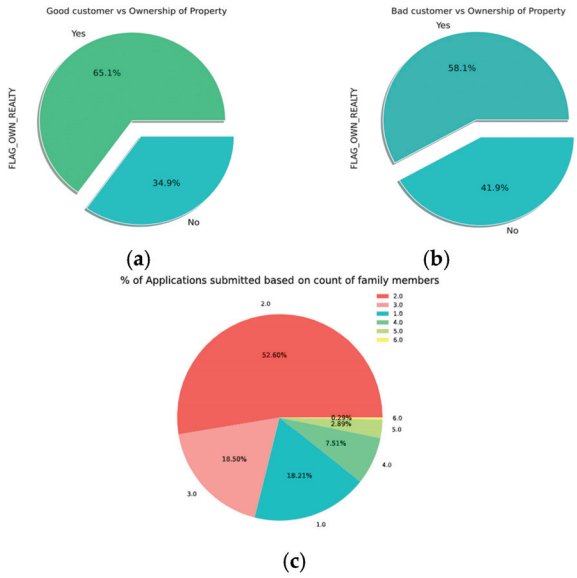


Figure 10. Pie charts of features. (**a**) Good customer vs. ownership of property; (**b**) bad customer vs. ownership of property; (**c**) applications vs. family members.

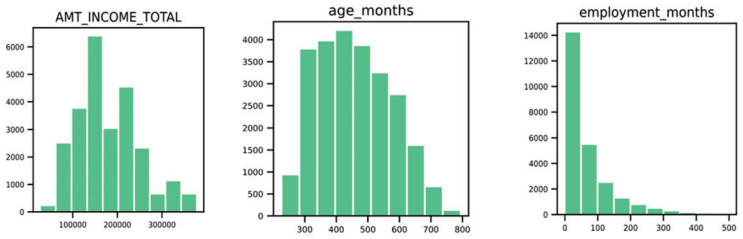


Figure 11. Histograms of features.

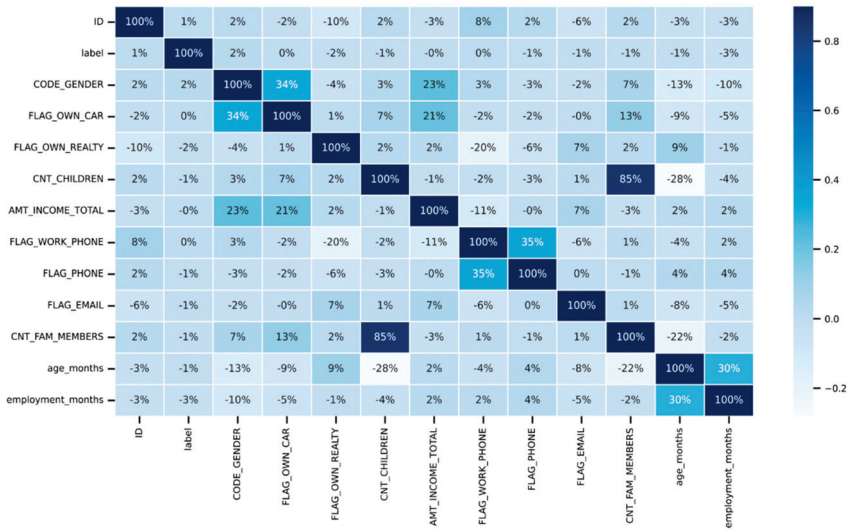


Figure 12. Correlation matrix.

3.3. Data Pre-Processing

One of the early pre-processing procedures that must be carried out in machine learning is splitting the modeling dataset into training and testing samples. Testing the model on the test set after training it on the training set will assist in determining how effectively the model generalizes to brand-new, untainted data. The training and testing sections of the dataset used in this study were split in a proportion of 70% training and 30% testing (70:30). Thirty percent of the data were utilized to assess the learning effectiveness of the testing data, while seventy percent of the datasets were utilized as a training set, ensuring that the training model was not exposed to the test sets.

When working with the classification model, the most significant problem is the unequal distribution of the values across the dataset. This is the fundamental issue. The same pattern is shown in the research data, where the distribution of good customers is 98.7% and bad customers are 1.3% (Figure 6). Most traditional machine-learning techniques assume that the distributions of the target classes are uniform. This impacts the models' performance because unbalanced datasets will cause these models to underperform. As a result, a higher accuracy for large categories (as in our case, good customers) and a lower accuracy for smaller classes (bad customers) may result from the direct input of unbalanced data. Even though the performance metrics show a good accuracy in this circumstance, other performance metrics will not show high enough ratings in other evaluation criteria. Two methods can resolve this issue. The first one is undersampling, which deletes the dominant values, and the second is oversampling, which adds rare values to the dataset. Undersampling strategies such as IB3, DROP3, and GA and oversampling strategies such as SMOTE, CTGAN, and TAN can be used to handle unbalanced data.

SMOTE is an oversampling method, one of the better ways of handling imbalanced data by producing fictitious samples for the minority class. This strategy assists in overcoming the issue of overfitting caused by random upsampling. It focuses on the feature space to generate new instances using interpolation between positive occurrences that are close together. The issue is handled and sorted by using the mentioned upsampling method, SMOTE. It involves adding artificially generated data records corresponding to the minority class into the dataset. After the upsampling procedure, the counts for the two labels are almost identical and can be used for predictive modeling.

According to our data, SMOTE can tackle the imbalanced target variance issue, and the testing data show that the distribution is likewise balanced, with 7355 "good customers" and 7473 "bad customers".

4. Machine-Learning Algorithms

The models presented below serve as illustrations of supervised machine-learning techniques utilized in the analysis. They are used to select the most appropriate classification model to classify good customers from bad customers, contributing to the credit card industry.

4.1. Random Forest

Prior to discussing the details of the random forest model, it is vital to define decision trees, ensemble models, and bootstrapping, all of which are fundamental to an understanding of the random forest model (Beheshti 2022). The decision tree is an application of supervised machine learning. It provides a flow chart type/tree-like structure, making it simple to visualize and extract information from the background process (Sharma 2023).

According to (Meltzer 2023), multiple decision trees are grown using random forests and combined to produce a more precise prediction, see Figure 13. The random forest algorithm develops trees while simultaneously introducing more randomness to the system. This method finds the best variable from a random collection of features when partitioning a node instead of selecting the element that is regarded as the most essential. This ends up providing vast diversity, which, in many circumstances, leads to a better model. As a

result, in a random forest, the method used to divide a node will only consider a random subset of all the characteristics (Donges 2021).

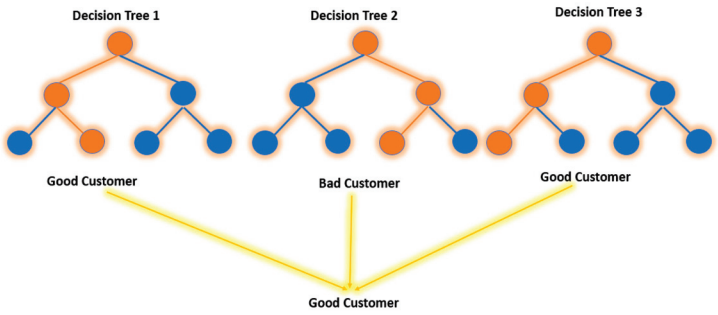


Figure 13. Decision tree.

The pseudo-code for the random forest model is shown in Algorithm 1 to elucidate the mechanism by which random forests operate. The algorithm describes a structured approach to random forests, which use bootstrap sampling to generate individual trees, which are then combined to form the overall forest.

Algorithm 1. Random Forest Algorithm

Precondition: A training set $X = \{(i_1, j_1), (i_2, j_2), \ldots, (i_n, j_n)\}$, features F, and number of trees in forest B.

1. Function RandomForest(X, F)
2. $H \leftarrow \emptyset$
3. For $i \in \{1, \ldots, B\}$ do
4. $X^i \leftarrow$ A bootstrap sample from S
5. $h_i \leftarrow \text{RandomizedTreeLearn}\left(X^i, F\right)$
6. $H \leftarrow H \cup$ (Sajumon 2015)
7. End for
8. Return H
9. End function
10. Function RandomizedTreeLearn(X, F)
11. At each node:
12. $f \leftarrow C \subset F$
13. Split on the best feature in f
14. Return the learned tree
15. End function

One of the hyperparameter tuning parameters is referred to as the "Criterion". The Gini index is set as the default parameter for the random forest. The Gini index is used to determine the node distribution on a decision tree branch.

$$Gini = 1 - \sum_{i=1}^{c} (p_i)^2$$

where

p_i = the probability of picking the data point with the class *i;*
c = number of classes.

4.2. Logistic Regression

Logistic regression is one of the supervised machine-learning algorithms classified as a subset of linear regression. However, it is solely employed for categorization. Logistic regression completes binary classification tasks by estimating the likelihood of a given

outcome, event, or observation. In our work, the model produces a binary outcome with only two options: good customer or bad customer, based on the status.

The pseudo-code for the logistic regression algorithm is shown in Algorithm 2. It splits data into different individual groups and examines the association between one or many independent features. It is frequently applied in the field of predictive modeling, in which the model determines the mathematical likelihood of whether or not a particular occurrence belongs to one specific category. As per our case, 0 is denoted as "good customer" and 1 is denoted as "bad customer". The sigmoid function is an activation function of logistic regression, and logistic regression employs the sigmoid function to establish a link between predictions and probabilities.

Algorithm 2. Logistic Regression

Logistic Regression

1. Input: Training data
2. Begin
3. For $i-1$ to k:
4. For each training data instance d_i:
5. Set target value for regression to $z_i = \frac{y_i - P(1|d_j)}{[P(1|d_j)(1-P(1|d_j))]}$
6. Initialize the weight of instance d_j to $[P(1|d_j)(1-P(1|d_j))]$
7. Finalize f(j) to data with class value z_j and weight w_j
8. Classical label decision
9. Assign (class label: 1) if $P_{id} > 0.5$, otherwise (class label: 2)
10. End

As shown in Algorithm 3, any real number can be transformed into a range of 0 and 1 using a function called the sigmoid, represented by an S-shaped curve. The sigmoid function's output is regarded as 1 if it is greater than 0.5. On the other side, the output is categorized as 0 if it is less than 0.5. The formula of the sigmoid function is as follows:

$$f(x) = \frac{1}{1+e^{-x}}$$

where e = base of the natural logarithms.

Algorithm 3. Sigmoid Function

Sigmoid Function
Require: Training data D, number of epochs e, learning rate η, standard deviation σ
Ensure: Weights $w_0, w_1, \ldots, w_k$

1. Initialize weights $w_0,\ w_1, \ldots,\ w_k$ from a standard normal distribution with zero mean and standard deviation σ
2. For epoch $1, \ldots, e$ do
3. For each $(x, y) \in D$ in random order do
4. $\hat{y} \leftarrow w_0 + \sum_{i=1}^{k} w_i x_i$
5. If $(\hat{y} > 1 \wedge y = 1) \cup (\hat{y} < 1 \wedge y = -1) \Rightarrow$ continue
6. $w_0 \leftarrow w_0 - \eta 2(\hat{y} - y)$
7. For i in $1, \ldots, k$ do
8. $w_i \leftarrow w_i - \eta 2(\hat{y} - y)x_i$
9. End For
10. End For
11. Return $w_0, w_1, \ldots, w_k$

Logistic regression is demonstrated by the following equation:

$$y = \frac{e^{(b_0 + b_1 x)}}{1 + e^{(b_0 + b_1 x)}}$$

where

x = input value;
y = predicted output;
b_0 = bias or intercept term;
b_1 = coefficient for input (x).

Parallel to linear regression, the above equation predicts the output value (0 or 1) by linearly integrating the input values using weights or coefficient values.

4.3. Neural Network

The model for predicting credit risk is developed using an artificial neural network, which has a structure with three layers called input nodes, hidden layers, and output layers, as illustrated in Figure 14. According to the dataset, customer details are the input nodes, and customer classifications such as "good customer" or "bad customer" are the output nodes.

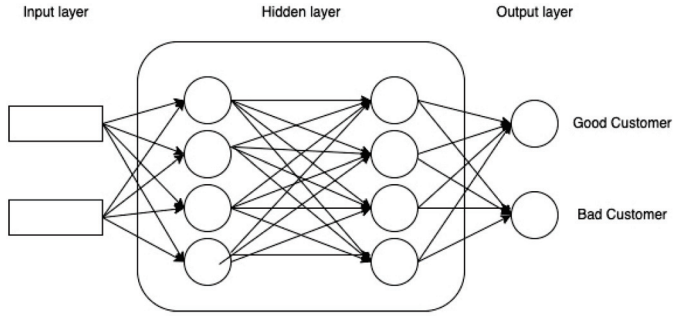


Figure 14. Neural network.

Training data are received by the input layer and passed to the hidden layers to convert the raw data into characteristics with a high dimension that are nonlinear, and then the output layer classifies the data. Firstly, the algorithm will be used to train the ANN-based classifier based on historical information. Later, the algorithm will be used to determine the customer's credit risk.

Generally, a neural network begins with a random set of weights. Each time the network finds an input–output pair, it modifies its weights based on it. Each pair passes through two processing phases: a forward pass and a backward pass. The forward pass involves delivering a representative input to the network and allowing activations to flow until they reach the output layer. The standard backpropagation is a gradient descent algorithm that repeats steps in the reverse direction to adjust the model's parameters based on weights and biases.

The algorithm's first iteration step can be expressed as follows:

$$W(t+1) = W(t) + \mu(-\nabla E(t))$$

where

$W(t)$ = vector of the weights at iteration step t;
$\nabla E(t)$ = current gradient of the error function;
E = sum of the squared errors;
μ = learning rate.

The learning model updates the gradient descent weights with more momentum (β) in order to shorten the training time.

$$W(t+1) = W(t) + \mu(-\nabla E(t)) + \beta \Delta W(t-1)$$
and
$$\Delta W(t) = \mu(-\nabla E(t)) + \beta \Delta W(t-1)$$

where

$\Delta W(t)$ = current adjustment of the weights;
$\Delta W(t-1)$ = previous change to the weights;
β = momentum.

Momentum enables a network to deploy to recent trends in the error surface and local gradient. It permits the network to dismiss small elements in error; it functions like a low-pass filter. Through the use of the first-order and also second-order derivatives of μ and β, the learning rates and momentum are updated with optimum rates during the training process. Each iteration of the backpropagation algorithm allows for the quick and easy computation of these derivatives. After the training, the model can be used to distinguish the riskiest customers by analyzing customer data.

To comprehend the intricate patterns that lie beneath the surface, deep-learning algorithms require an adequate quantity of data. When more data are used, the performance of deep-learning models will significantly improve.

4.4. XGBoost Algorithm

XGBoost (see Figure 15) is a widely used implementation of the gradient-boosted tree technique that is both efficient and open-source. Gradient boosting is a method of supervised learning that accurately predicts a target variable by combining the predictions of a series of weaker, simpler models.

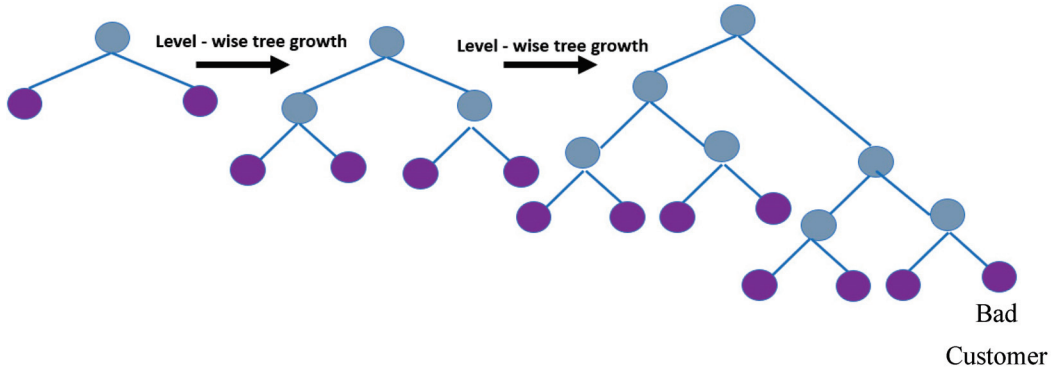


Figure 15. XGBoost.

According to (Kharwal 2020), gradient boosting works well by minimizing loss when adding new models. Regression trees serve as the weak learners in gradient boosting for regression to translate each input parameter to a leaf that has a continuous value. XGBoost minimizes a systemized objective function by integrating the convex algorithm based on the variance between the anticipated and target outcomes and a penalty element for model complexity. The training procedure is carried out repeatedly by adding additional trees that reveal the residuals or mistakes of older trees, which are then incorporated with earlier trees to produce the final forecast.

$$F_m(X) = F_{m-1}(X) + \alpha_m h_m(X, r_{m-1})$$

where

α_i—regularization parameters;
r—residuals computed with the *i*th tree;
h_i—function trained to predict residuals;
r_i—using X for the *i*th tree.
Residuals and $\text{Arg}(\min_\alpha)$ have to be computed in order to compute the α.

where

$$\text{Arg}(\min_\alpha) = \sum_{i=1}^{m} L(Yi, Fi-1(Xi) + \alpha hi(Xi, ri-1))$$

where

L(Y, F(X)) is the differentiable loss function.

4.5. LightGBM Algorithm

LightGBM represents a gradient-boosting framework created on decision trees (a tree-based algorithm that grows leaf-wise, as shown in Figure 16, rather than level-wise) for the light gradient to reduce memory usage, enhance memory utilization, and increase the model's efficiency.

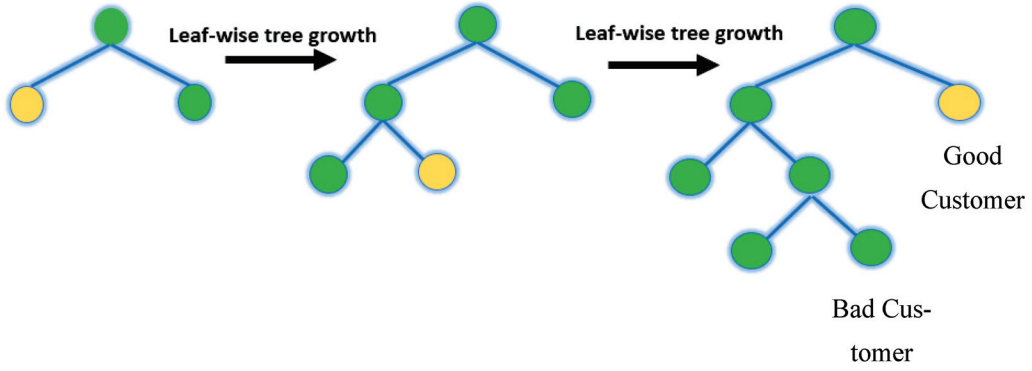


Figure 16. LightGBM.

As per (GeeksforGeeks n.d), LightGBM employs two methods: EFB (Exclusive Feature Bundling) and GOSS (Gradient-based One-Side Sampling), which collectively enable the model to function effectively. GOSS will merely use the remaining data to evaluate the overall information gain, excluding the extensive number of data sections that have insignificant gradients. The calculation of statistics expansion gives more weight to the data instances with significant gradients. Despite utilizing a smaller dataset, GOSS can produce trustworthy findings with substantial information gain compared to other models. It has gained popularity due to its high speed and ability to handle large amounts of data with low space complexity. Although EFB rarely receives any non-zero values parallel to shrinking the number of characters, it does put the mutually exclusive features along with frivolity. This affects the total outcome for efficient feature elimination without compromising the split point's accuracy. Any algorithm's training time will be shortened by 20 times by combining the two improvements. With EFB and GOSS together, LGBM can be considered gradient-enhancing trees. It performs best with massive data.

LightGBM and XGBoost vary primarily in that whereas XGBoost employs a histogram-based method and a pre-sorted algorithm for the most effective division calculation, LightGBM selects data instances to calculate a split value using the GOSS technique. LightGBM employs a highly optimized decision-making algorithm that is based on histograms and offers significant advantages while also being efficient and memory-efficient.

The critical characteristics of LGBM include higher accuracy and faster training speeds, low memory usage, superior comparative accuracy to other boosting algorithms, better handling of overfitting when working with smaller datasets, support for parallel learning, and compatibility with small and large datasets.

Decision tree-based machine-learning algorithms were formerly the industry standard. The best solutions for the majority of problems used XGBoost. Microsoft unveiled its gradient-boosting technology, LightGBM, a few years ago. Currently, it takes center stage in gradient-boosting devices. XGBoost has been replaced by LightGBM.

4.6. AdaBoost Algorithm

A common boosting approach called AdaBoost aids in combining several "weak classifiers" into one "strong classifier". In other words, to improve weak classifiers and make them stronger, AdaBoost employs an iterative process. AdaBoost is a form of

ensemble learning approach. Based on the output of the previous classifier, it aids in selecting the training set for each new classifier. It establishes how much weight must be assigned to each classifier's suggested response when the results are combined.

Algorithm 4 clarifies the AdaBoost methodology. Initially, all data points are equally weighted. However, as the algorithm progresses through each iteration, it meticulously recalibrates the weights of incorrectly classified data points. This weight adjustment ensures that subsequent classifiers prioritize those specific misclassified instances, thereby improving the cumulative prediction accuracy. At the end of the iterative process, AdaBoost merges the outcomes of all weak classifiers and weights them according to their respective accuracies to produce a robust final classifier.

Algorithm 4. AdaBoost

AdaBoost Algorithm
Given: $(x_1; y_1), \ldots, (x_m; y_m), x_i \in X,\ y_i \in Y = \{-1, 1\}$.
Initialize: $D_1(i) = \frac{1}{m}$.
For $t = 1, ..., T$:

1. Train a weak classifier using distribution D_t.
2. Obtain the weak hypothesis $h_t : X \to \{-1, 1\}$ with the error:

$$\varepsilon_t = \sum_{i:h_t(x_i) \neq y_i} D_t(X_i)$$

3. Choose $\alpha_t = \frac{1}{2}\log\frac{1-\varepsilon_t}{\varepsilon_t}$.
4. Update:

$D_{t+1}(i) = \frac{D_i(i)}{Z_t} = e^{-\alpha_t}$ if i is correctly classified, otherwise e^{α_t}

where Z_t is a normalization factor chosen such that $\sum_{i=1}^{m} D_{t+1}(i) = 1$

5. Output: The final hypothesis:

$$H(x) = \sum_{t=1}^{T} \alpha_t h_t(x)$$

The research by (Nazarenko et al. 2019) highlighted the benefits of AdaBoost:

- **Good generalization skills:** Creating compositions for real-world issues of a higher caliber is more feasible than using the fundamental algorithms. As the number of fundamental algorithms rises, the generalization ability may become more effective (in some missions).
- **Own boosting expenses are minimal:** The training time of the fundamental algorithms nearly entirely determines the amount of time needed to construct the final image.
- Several drawbacks of AdaBoost can be discussed as follows:
- **Boosting technology develops gradually:** It is crucial to guarantee high-quality data.
- **Tactful to unorganized data and outliers**: Hence, it is intensely advised to avoid these before using AdaBoost.
- **Slower than XGBoost:** It has also been indicated that AdaBoost is slower than XGBoost.

Hyperparameter tuning aims to train each model and finalize the best-predicting model. In the modeling process, one critical aspect that needs to be considered is whether there is still potential for improvement before completing the best-performing model. As a result, the model needs to be improved in whatever manner possible. Hyperparameters are one of the key elements in performance improvement. The performance of these models can be considerably improved by selecting the appropriate values for their hyperparameters, which are critical to how effectively they function. Table 2 represents the hyperparameters set for each model for the predictions.

Table 2. Hyperparameter for each methodology.

Methodology	Hyperparameter	Value
Random Forest	n_estimators	250
	min_samples_leaf	16
	maximum depth	12
	random State	42
XGBoost	max_depth	12
	n_estimators	250
	min_child_weight	8
	subsample	0.8
	learning_rate	0.02
	seed	60
	gamma	0
	colsample_bytree	0.8
	objective	binary: logistic
LGBM	num_leaves	50
	learning_rate	0.02
	n_estimators	250
	subsample	0.8
	colsample_bytree	0.8
AdaBoost	max_depth	1
	n_estimators	100
	learning_rate	1
Neural Network	hidden_layer_sizes	400,800
	max_iter	1000
	random_state	25
	shuffle	TRUE
Logistic Regression	random_state	42

5. Implementation and Results

In this section, we first introduce confusion matrix, which is the framework that we adopted; and then we will discuss the results from implementing our chosen machine-learning algorithms. Further analysis will also be investigated from the best-performing model.

5.1. Confusion Matrix

The confusion matrix, shown in Figure 17, is a technique for determining the effectiveness of a classification algorithm. Classification accuracy is defined as the proportion of accurate predictions compared to all other predictions. A better understanding can be obtained of what the classification model is doing correctly and the mistakes it makes by calculating a confusion matrix. The total number of precise predictions for a class is recorded in both the predicted column and true label row for that class value. Likewise, the total amount of imprecise predictions for a class is recorded in both the predicted column and the actual label row for that class value. Confusion matrices attempt to differentiate the occurrences with a specific outcome.

The binary classification problem distinguishes between observations with a particular outcome and regular observations. Based on our model predictions, the customers will become default or not in the loan defaulting prediction. The good customer is labeled as "0" and the bad customer as "1" in this matrix.

This results in the following:

- True positive—When good customers are accurately predicted as good customers.
- False positive—When bad customers are improperly predicted as good customers.
- True negative—When bad customers are accurately predicted are bad customers.
- False negative—When good customers are improperly predicted as bad customers.

The equations to compute the respective rates are as follows:

- True positive = Number of customers accurately predicted as good/Actual number of good customers
- False positive = Number of customers improperly predicted as good/Actual number of bad customers
- True negative = Number of customers accurately predicted as bad/Actual number of bad customers
- False negative = Number of customers improperly predicted as bad/Actual number of good customers

According to the test samples, 7422 are good customers, whereas 7406 are bad customers. Based on the number of actual good and bad customers for the prediction algorithm, we will analyze the false positive rate, true positive rate, false negative rate, and true negative rate for the best-performing and worst-performing algorithms.

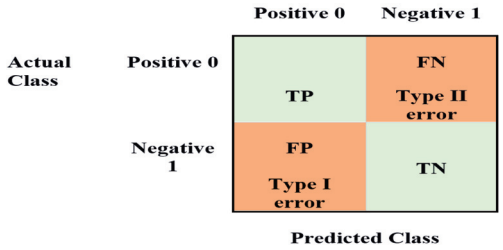


Figure 17. Confusion matrix.

5.2. Implementation and Comparison of Machine-Learning Algorithms

The neural network model forecasted all customers as good, as shown in Figure 18. As a result, a total of 5918 bad customers were inaccurately predicted as good customers. In addition, none of the bad customers were correctly predicted as bad customers by the neural network. Consequently, the false positive rate ended up being 80%, whereas the true negative rate was 20%.

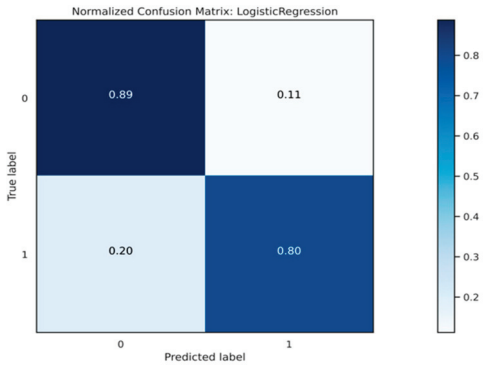


Figure 18. Confusion matrix—Logistic regression.

Since the model projected that nearly all consumers would be good, the true positive rate is calculated as 89%. Moreover, the model incorrectly projected 0 customers as bad customers when they were actually bad customers, resulting in a false negative rate of 20%.

The XGB model correctly predicted both 7406 good customers and 7320 bad customers, as shown in Figure 19. Consequently, the true positive rate ended at 99.50%, whereas the true negative rate ended at 96.27%. In addition, the model incorrectly projected 278 customers as good customers when they were bad customers. Additionally, the model incorrectly predicted that 37 customers were bad when they were good customers.

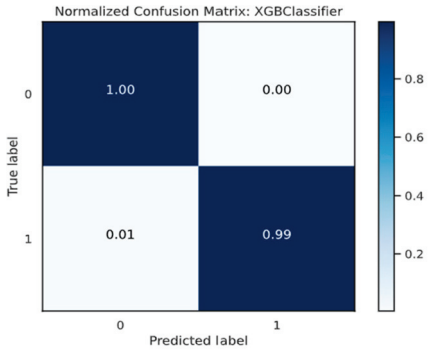


Figure 19. Confusion matrix—XGBoost.

As per the confusion matrix, XGBoost displays a superior performance, while the neural network does not perform well for this purpose. Other performance metrics are also vital to know how the model performs in the prediction scenario. This examination focuses on classifier evaluation metrics, including AUC, accuracy, recall, and precision.

5.2.1. Accuracy

Accuracy is the easiest and most well-known measure for classification problems. It is calculated by dividing the number of accurate predictions by the overall number of forecasts. Further, while discussing accuracy, true negative rate (TNR), true positive rate (TPR), false negative rate (FNR), and false positive rate (FPR) also need to be considered. Accuracy can be computed as follows:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

As shown in Figure 20, the bar chart of accuracy comparison reveals that XGB has the highest accuracy, coming in at 99.4%, followed by LGBM, which has a good accuracy of 99.3%, and finally, logistic regression, which has the lowest accuracy, coming in at 84.3%.

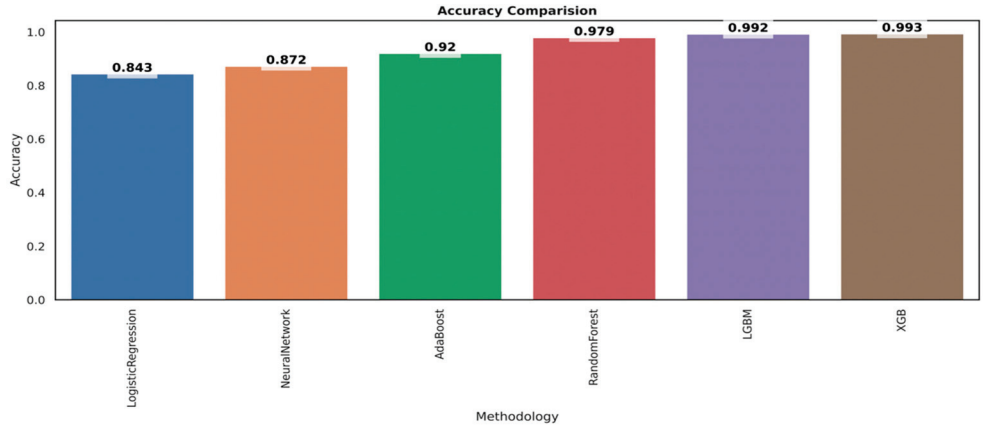


Figure 20. Performance evaluation.

However, accuracy alone does not tell the entire story when dealing with a class-unbalanced dataset, where there is a considerable variance between the total number of positive and negative labels. Therefore, other performance metrics are also further analyzed.

5.2.2. Recall and Precision

Precision and recall apply to individual classes only; for instance, recall for good customers or precision for bad customers.

Precision attempts to answer the question, "What percentage of positive identifications were actually accurate?" The basis for precision is prediction. To explain, how many were correctly predicted as bad customers out of all the bad customer predictions? Or how many were correctly predicted as good customers out of all the good customers' predictions?

The mathematical equation for precision is as follows:

$$\text{Precision} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Positive}}$$

Recall attempts to answer the question, "What percentage of true positives were correctly detected?" The basis of the recall is the truth. That is, out of all the bad customers, how many were predicted as actually bad customers? Or, out of all the good customers, how many were predicted as actually good customers?

The mathematical equation for the recall is as follows:

$$\text{Recall} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Negative}}$$

Precision or recall should be chosen depending on the problem that needs to be solved. Use precision if the issue is sensitive to classifying a sample as positive in general, including negative samples that were mistakenly categorized as positive. Use recall if the objective is to find every positive sample without minding whether some negative samples could be mistakenly categorized as positive. In our situation, identifying bad customers is very similar to identifying good customers. According to this, precision is critical in our scenario.

The precision and recall comparison bar charts in Figures 21 and 22, respectively, make it evident that the model XGB has the highest possible precision as per Figure 21 and recall as per Figure 22 (both are 0.994), followed by the model LGBM (both are 0.992 and 0.993). In addition, the logistic regression has the lowest score for both precision (0.846) and recall (0.843).

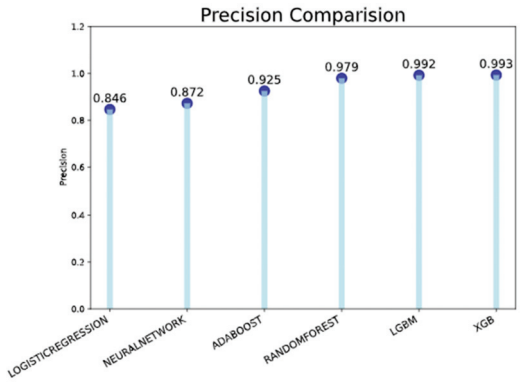


Figure 21. Precision comparison.

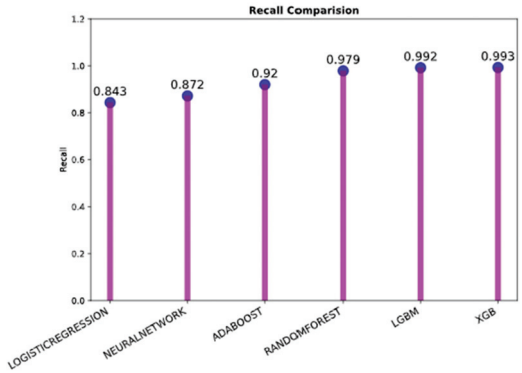


Figure 22. Recall comparison.

5.2.3. F1 Score

Another performance metric is the F1 score, which ranges from 0 to 1 and is a harmonic average of recall and precision. The most recommended quality metric for a binary classification task is to optimize for its F1 score. The higher the F1 score, with 0 being the worst and 1 being the highest, the better the overall performance of the model. Only when precision and recall are both 100% does it attain its ideal level of 1. The F1 score has its worst value of 0 if one of them is equal to 0.

$$F1 = 2 \times ((\text{Precision} \times \text{Recall})/(\text{Precision} + \text{Recall}))$$

The scenario is reflected in Figure 23 in the same way as recall and precision discussed earlier. It is clear from this that the model XGB achieves the highest possible F1 score due to the fact that the XGB achieves the highest possible recall and precision.

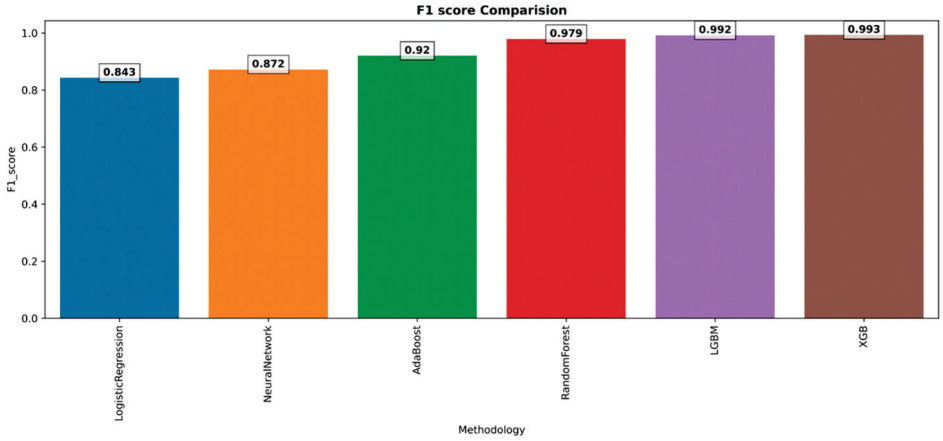


Figure 23. F1 score.

5.2.4. ROC Curve and AUC Score

The ROC curve and AUC score are among the most crucial evaluation systems for measuring the performance of classification models. The ROC curve is a probability curve, and the AUC defines the degree or measure of separability. It describes how well the model differentiates between classes. The greater the AUC, the better the model predicts bad credit card customers as bad and good credit card customers as good. In summary, the greater the AUC, the better the model differentiates between good and bad credit card customers. The following two factors make AUC desirable:

- **Scale is unimportant to AUC:** It evaluates how well predictions are ranked rather than the absolute values of the predictions.
- **AUC is independent of the classification threshold:** It evaluates how well the model predicts regardless of the classification threshold.

The baseline of the ROC curve can be explained at the diagonal points by default (FPR is equal to TPR). TPR is mapped against FPR on the ROC graph, with FPR on the x-axis and TPR on the y-axis. Algorithms closer to the top left corner corresponding to the coordinate (0, 1) in the Cartesian plane demonstrate a better performance than those below.

The test will be less precise the closer the graph gets to the ROC plot's 45-degree diagonal. The fact that the ROC curve does not depend on the class distribution is one of the many reasons it is so valuable. It enables and facilitates situations in which the classifiers predict unusual events, which is the same as our concern regarding the detection of bad customers.

The value of AUC ranges from 0 to 1. An AUC of 0 specifies a model with 100% incorrect predictions, while an AUC of 1 indicates a model with 100% correct predictions. If the area under the curve AUC is equivalent to 0.5, then we can conclude that the algorithm is incapable of differentiating between good customers and bad customers accordingly.

On an ROC curve, a greater value on the *x*-axis specifies a more significant number of false positives than true negatives. At the same time, a higher value on the *y*-axis also represents a more significant proportion of true positives than false negatives. Accordingly, threshold selection depends on the capacity to create an equilibrium between false positives and negatives. Model comparison of random forest, neural networks, XGB, LGBM, AdaBoost, and logistic regression is as follows.

As per Figure 24, LGBM and XGB show a better performance. The best models for correctly classifying observations are LGBM and XGB, which have the most significant AUC and the highest space below the curve. The green line indicates the model LGBM and is embedded behind the red line model XGB; after XGB and LGBM, random forest and AdaBoost perform best—in that order. Logistic regression is near the points lying around the diagonal, and the neural network is on the diagonal line, which indicates a poor performance.

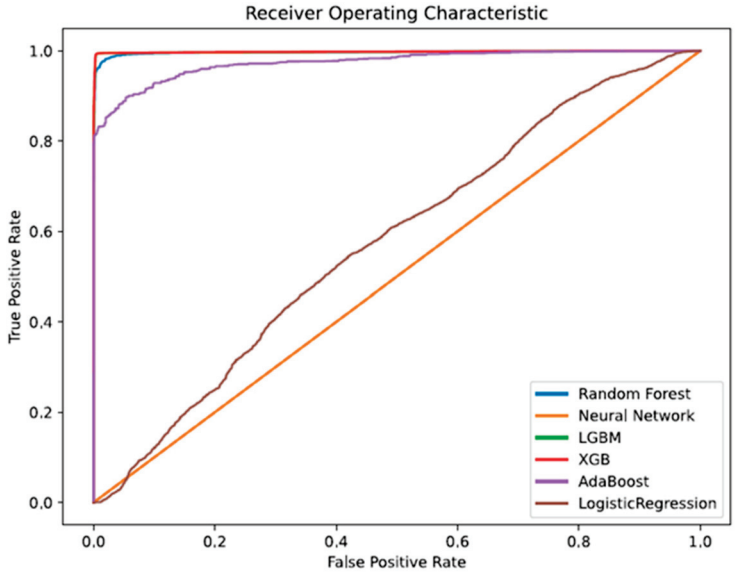


Figure 24. ROC–AUC curve.

As per Figure 25 below, the red line of XGBoost at the (0, 1) position in the upper left corner of the Cartesian plane exhibits a superior performance.

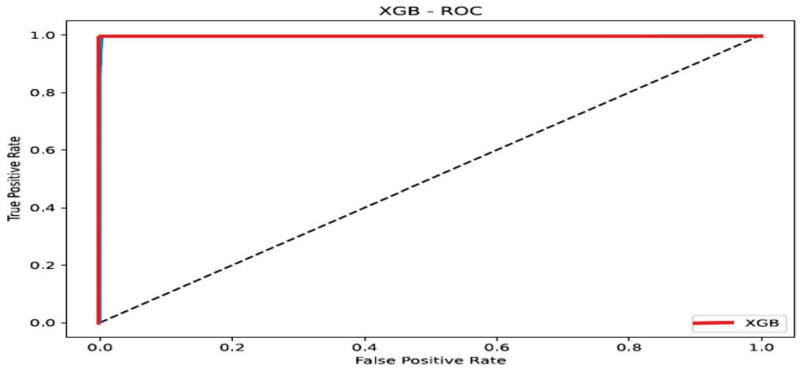


Figure 25. ROC—XGB.

As a result of our analysis of ROC–AUC, we are able to draw the conclusion that LGBM and XGB outperform the other algorithms in terms of their ability to identify good customers as good and bad customers as bad.

5.2.5. MCC

A class imbalance can affect accuracy, recall, precision, and F1 score, making them all uneven. An alternative approach to binary classification is to treat the true class and the predicted class as two different variables and calculate their correlation coefficient similarly to compute the correlation coefficient between any two variables. MCC aids in identifying the classifier's shortcomings, particularly with regard to the negative class samples. MCC is a single-value statistic that distills the confusion matrix, much like the F1 score. No class is more significant than any other since MCC is also totally symmetric; even when the positive and negative values are switched, the result remains the same. A high number, near 1, indicates that MCC correctly predicts both classes. In other words, a score of 1 represents complete agreement. Even if one class is unreasonably under-represented or over-represented, MCC considers all four values in the confusion matrix. Following this, there are calculations for MCC:

$$\text{MCC} = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

By looking at their equations, one can quickly determine the main advantage of employing MCC instead of the F1 score. The number of true negatives is ignored by the F1 score. Conversely, MCC is gracious enough to take care of all four entries in the confusion matrix. MCC is favored over the F1 score only if the cost of low precision and low recall is truly unknown or unquantifiable because it is a "fairer" evaluation of classifiers, regardless of which class is positive.

MCC is the best single-value classification metric, which serves to summarize the confusion matrix or an error matrix. As per Figure 26, the model XGB outperforms the other algorithms with a score of 0.9879, followed by LGBM (0.986).

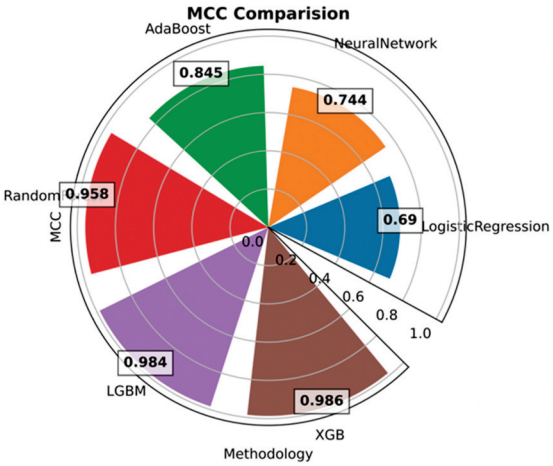


Figure 26. MCC evaluation.

5.3. Comparison of the Best- and Worst-Performing Models

According to Figure 27, random forest is superior in its ability to differentiate between good and bad customers. The best scores assure the breakdown results for both good customers (0) and bad customers (1) with the other performance metrics, such as precision (0–0.965 and 1–0.995), recall (0–0.995 and 1–0.964), and f1 score (0–0.980 and 1–0.979).

The following Table 3 represents the summary of all performance metrics for each algorithm.

Finally, we look at the feature importance from the best-performing model XGBoost, and the results are shown in Figure 28 below, which indicates how each feature contributes to the classification prediction model of XGBoost.

```
Accuracy Score is 99.319
      0     1
0  7398    24
1    77  7329
[[0.99676637 0.00323363]
 [0.01039698 0.98960302]]
              precision    recall  f1-score   support

           0      0.990     0.997     0.993      7422
           1      0.997     0.990     0.993      7406

    accuracy                          0.993     14828
   macro avg      0.993     0.993     0.993     14828
weighted avg      0.993     0.993     0.993     14828

Overall Accuracy 99.3
Overall Precision 99.3
Overall Recall  99.3
Overall f1  99.3
Matthew's Correlation Coefficient: 0.986
```

Figure 27. Summary of the best-performing model.

Table 3. Summary of performance metrics.

Algorithm	Accuracy	Precision	Recall	F1 Score	ROC–AUC	MCC
Random Forest	97.9%	0.979	0.979	0.979	0.996	0.958
Neural Network	87.2%	0.872	0.872	0.872	0.942	0.744
XGB	**99.3%**	**0.993**	**0.993**	**0.993**	**0.997**	**0.986**
LGBM	99.2%	0.992	0.992	**0.993**	**0.997**	0.744
AdaBoost	0.920	0.925	0.920	0.920	0.976	0.845
Logistic Regression	0.843	0.846	0.843	0.843	0.910	0.690

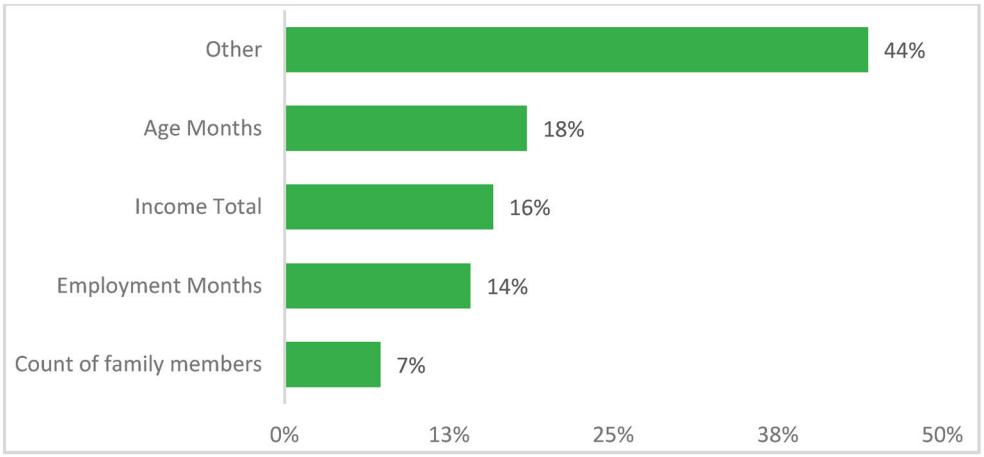


Figure 28. Feature importance of XGBoost.

A feature with a higher value indicates that it is more crucial than a feature with a lower value in predicting the customer type. In addition, the inclusion/exclusion of this feature from the training set significantly impacts the final results. As per the figure above, the customer's age, income total, employment months (experience), and count of family members are the crucial features of the XGBoost model formation. Also, other features only contribute around 13% to the model.

6. Further Discussion

From the experimental results in this section, the XGBoost model demonstrated high accuracy on the selected dataset; this is in line with the recent literature on credit risk modeling, which emphasizes the importance of dynamically robust machine-learning models that maintain performance over time. The work of (Shi et al. 2022) discusses various computing techniques, including traditional statistical learning, machine learning, and deep learning, highlighting how these models have evolved to meet the demands of credit risk prediction. The authors underscore the necessity for the continuous adaptation of these models to new data and market changes, reinforcing the idea that model robustness is critical for real-world financial applications. Similarly, research by (Alonso Robisco and Carbó Martínez 2022) focuses on the model risk-adjusted performance of machine-learning algorithms in credit default prediction, identifying the potential risks when models overly depend on specific datasets. The study explores the application of interpretability techniques, such as SHAP and LIME, to ensure that models can be regularly evaluated and validated against changing market conditions. It also proposes methodologies to quantify model risk, underscoring the need for dynamic evaluation processes to ensure models remain effective and compliant with regulatory standards.

It is crucial to recognize that credit risk profiles evolve over time due to various factors, including economic fluctuations and changing consumer behaviors. To maintain accuracy and robustness, models should be periodically retrained with new data to adapt to these changes.

7. Conclusions

7.1. Summary

Collecting payments from bad customers is a significant challenge for banks and financial institutions, leading to prolonged collection processes and high expenses. One of the most crucial decisions for these organizations is accurately identifying customers who are both willing and capable of repaying their debts. This research contributes to this decision-making process by developing a model to predict the credit risk of credit card customers.

By addressing outliers and selecting essential features, various machine-learning models were applied to a credit dataset. Among the models examined, XGBoost outperformed others in terms of all performance metrics, including accuracy, precision, recall, ROC–AUC, F1 score, and MCC.

The proposed XGBoost model can effectively predict the default status of credit card applicants, aiding banks and financial institutions in making more informed decisions. By implementing this model, banks can enhance the accuracy of their credit risk assessments, resulting in higher acceptance rates, increased revenue, and reduced capital loss. This approach allows financial institutions to operate efficiently, maximizing profits while minimizing costs.

For banks and financial regulators, this research has real advantages. Lenders can improve their credit approval procedures and possibly lower default rates and related losses by putting the XGBoost model into practice. For example, over a twelve-month period, a large bank that tested a comparable methodology experienced a 15% drop in defaults. These insights could be used by regulators to revise credit risk assessment rules, guaranteeing that banks continue to employ sound evaluation techniques. Machine-learning models are already being considered by one regulatory authority as a requirement for standard credit checks. Additionally, the model's capacity to pinpoint important risk indicators may aid in the development of more focused financial education initiatives that target particular default risk-causing behaviors.

7.2. Recommendations

In light of the insights obtained from this study, it is recommended that credit card issuers or banks incorporate this predictive credit risk analytics into their operations for

strategic business and marketing decisions. Firstly, this method segments cardholders based on credit risk, allowing credit card issuers or banks to enhance their credit risk management by avoiding or decreasing the number of non-payers. Secondly, with a clear understanding of all the customers' creditworthiness, credit card issuers or banks can identify those most likely to responsibly address increased credit lines or additional financial products. This can lead to effective upselling and cross-selling strategies, thus increasing the customer lifetime value and overall revenue growth. Thirdly, credit card issuers or banks can implement a risk-based pricing strategy by accurately classifying card holders according to their credit risk. This may involve offering better terms to low-risk customers, enhancing customer retention, and attracting other creditworthy customers. Fourthly, economic fluctuations can affect the financial stability of cardholders and, therefore, their credit risk. The predictive credit risk analytics can quickly incorporate these changes by monitoring a customer's financial situation over a period of six months or even longer. This enables card issuers or banks to adjust their strategies accordingly, demonstrating resilience and flexibility in volatile market conditions.

Our research findings have pragmatic and practical implications for credit card companies, banks, and financial institutions, significantly improving lending decisions by leveraging the ML models, such as LGBM and XGBoost, to predict credit risk. Financial institutions can identify high-risk applicants to minimize bad debt and write-offs. Similarly, they can approve creditworthy customers to expand the market share. Increased predictive power also enables personalized loan terms based on an applicant's true default risk.

7.3. Limitations of the Study and Future Work

This research could be enhanced by incorporating more recent and diverse credit card data, which should include a wider range of features. Additionally, considering the acquisition of a dataset with a larger number of records would improve the robustness of the model.

The SVM algorithm was also attempted for prediction. However, it was excluded from this study because it required a significant amount of time (more than 2 h) to run, making it an expensive model in terms of computational resources.

The default credit status is influenced by macroeconomic factors such as the inflation rate, interest rate, GDP (gross domestic product), and unemployment rate. Future data collection efforts should incorporate these macroeconomic factors to enhance the prediction of credit card default status.

In subsequent work, the K-fold validation methodology should be incorporated, in addition to the machine-learning classifiers used in this study. This would add a new dimension to the prediction method, employing the latest techniques and expandable AI to better identify customers with good credit records.

Future work should also explore adaptive learning methods, such as online learning, that can dynamically adjust model parameters in response to new data patterns. Additionally, continuous performance monitoring in real-world applications is recommended to promptly identify and address any declines in model accuracy.

Author Contributions: Conceptualization, V.C. and S.S.; methodology, V.C.; software, S.S.; validation, V.C., H.W., and S.T.W.; formal analysis, V.C., S.S., and H.W.; investigation, V.C. and S.S.; resources, V.C.; data curation, S.S.; writing—original draft preparation, V.C. and S.S.; writing—review and editing, V.C., H.W., S.T.W., and J.L.; visualization, V.C., S.S., and M.A.G.; supervision, V.C. and M.A.G.; project administration, V.C.; funding acquisition, V.C. All authors have read and agreed to the published version of the manuscript.

Funding: This research is partly funded by VC Research [(VCR 0000209)] for Victor Chang.

Data Availability Statement: Data are available on https://www.kaggle.com/code/rikdifos/eda-vintage-analysis/data, accessed on 1 July 2024.

Acknowledgments: Authors are grateful to Qianwen Ariel Xu and Karl Drazar Hall, who helped to improve the quality of the paper in the early stage.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A

The explanation and remarks are obtained from https://www.kaggle.com/code/rikdifos/eda-vintage-analysis/data (accessed on 1 July 2024) for both datasets.

References

Adha, Risfian, Siti Nurrohmah, and Sarini Abdullah. 2018. Multinomial Logistic Regression and Spline Regression for Credit Risk Modelling. *Journal of Physics: Conference Series* 1108: 012019. [CrossRef]

Ali, Mohsin, Abdul Razaque, Joon Yoo, Uskenbayeva R. Kabievna, Aiman Moldagulova, Satybaldiyeva Ryskhan, Kalpeyeva Zhuldyz, and Aizhan Kassymova. 2024. Designing an Intelligent Scoring System for Crediting Manufacturers and Importers of Goods in Industry 4.0. *Logistics* 8: 33. [CrossRef]

Alonso Robisco, Andrés, and José Manuel Carbó Martínez. 2022. Measuring the model risk-adjusted performance of machine learning algorithms in credit default prediction. *Financial Innovation* 8: 70. [CrossRef]

Al-qerem, Ahmad, Ghazi Al-Naymat, and Mays Alhasan. 2019. Loan Default Prediction Model Improvement through Comprehensive Preprocessing and Features Selection. Paper presented at the 2019 International Arab Conference on Information Technology (ACIT), Al Ain, United Arab Emirates, December 3–5.

Aswini, Ravichandran, Balamurugan Muruganantham, Santosh Kumar, and Arumugam Murugan. 2020. Exploratory Data Analysis for Social Big Data Using Regression and Recurrent Neural Networks. *Webology* 17: 922–36. [CrossRef]

Bao, Wang, Ning Lianju, and Kong Yue. 2019. Integration of unsupervised and supervised machine learning algorithms for credit risk assessment. *Expert Systems with Applications* 128: 301–15. [CrossRef]

Behera, Sudersan, Sarat Chandra Nayak, and Annadanapu Pavan Kumar. 2023. A Comprehensive Survey on Higher Order Neural Networks and Evolutionary Optimization Learning Algorithms in Financial Time Series Forecasting. *Archives of Computational Methods in Engineering* 30: 4401–48. [CrossRef]

Beheshti, Nima. 2022. Random Forest Classification. Available online: https://towardsdatascience.com/random-forest-classification-678e551462f5 (accessed on 31 October 2024).

Bindal, Anirudh, and Sandeep Chaurasia. 2018. Predictive Risk Analysis For Loan Repayment of Credit Card Clients. Paper presented at the 2018 3rd IEEE International Conference on Recent Trends in Electronics, Information & Communication Technology (RTEICT), Bangalore, India, May 18–19.

Buchanan, Bonnie G., and Danika Wright. 2021. The impact of machine learning on UK financial services. *Oxford Review of Economic Policy* 37: 537–63. [CrossRef]

Chang, Victor, Raul Valverde, Muthu Ramachandran, and Chung-Sheng Li. 2020. Toward business integrity modeling and analysis framework for risk measurement and analysis. *Applied Sciences* 10: 3145. [CrossRef]

Chen, Tianqi, and Carlos Guestrin. 2016. XGBoost: A Scalable Tree Boosting System. Paper presented at the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13–17; pp. 785–94.

Chen, Xihui, Xuyuan You, and Victor Chang. 2021. FinTech and commercial banks' performance in China: A leap forward or survival of the fittest? *Technological Forecasting and Social Change* 166: 120645. [CrossRef]

Dm, Obare, and Muraya Mm. 2018. Comparison of Accuracy of Support Vector Machine Model and Logistic Regression Model in Predicting Individual Loan Defaults. *American Journal of Applied Mathematics and Statistics* 6: 266–71.

Donges, Niklas. 2021. A Complete Guide to the Random Forest Algorithm. Available online: https://builtin.com/data-science/random-forest-algorithm (accessed on 31 October 2024).

Duan, Jing. 2019. Financial system modeling using deep neural networks (DNNs) for effective risk assessment and prediction. *Journal of the Franklin Institute* 356: 4716–31. [CrossRef]

Education, IBM Cloud. 2020. What Is Exploratory Data Analysis (EDA)? Available online: https://www.ibm.com/topics/exploratory-data-analysis (accessed on 31 October 2024).

GeeksforGeeks. n.d. *LightGBM (Light Gradient Boosting Machine)*. Available online: https://www.geeksforgeeks.org/lightgbm-light-gradient-boosting-machine/ (accessed on 31 October 2024).

Guégan, Dominique, and Bertrand Hassani. 2018. Regulatory learning: How to supervise machine learning models? An application to credit scoring. *The Journal of Finance and Data Science* 4: 157–71. [CrossRef]

Han, Jiawei, Micheline Kamber, and Jian Pei. 2012. *Data Mining Concepts and Techniques*, 3rd ed. Waltham: Morgan Kaufmann Publishers.

Kharwal, Aman. 2020. Boosting Algorithms in Machine Learning. Available online: https://thecleverprogrammer.com/2020/10/30/boosting-algorithms-in-machine-learning/ (accessed on 31 October 2024).

Krawczyk, Bartosz. 2016. Learning from imbalanced data: Open challenges and future directions. *Progress in Artificial Intelligence* 5: 221–32.

Lin, Cian, Chih-Fong Tsai, and Wei-Chao Lin. 2023. Towards hybrid over- and under-sampling combination methods for class imbalanced datasets: An experimental study. *Artificial Intelligence Review* 56: 845–63. [CrossRef]

Liu, Lulu. 2022. A Self-Learning BP Neural Network Assessment Algorithm for Credit Risk of Commercial Bank. *Wireless Communications and Mobile Computing* 2022: 9650934. [CrossRef]

Lucarelli, Giorgio, and Matteo Borrotti. 2020. A deep Q-learning portfolio management framework for the cryptocurrency market. *Neural Computing and Applications* 32: 17229–44. [CrossRef]

Lundberg, Scott M., and Su-In Lee. 2017. A unified approach to interpreting model predictions. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*. Long Beach: Curran Associates Inc., pp. 4768–77.

Ma, Xiaojun, Jinglan Sha, Dehua Wang, Yuanbo Yu, Qian Yang, and Xueqi Niu. 2018. Study on a prediction of P2P network loan default based on the machine learning LightGBM and XGboost algorithms according to different high dimensional data cleaning. *Electronic Commerce Research and Applications* 31: 24–39. [CrossRef]

Maldonado, Sebastián, Juan Pérez, and Cristián Bravo. 2017. Cost-based feature selection for Support Vector Machines: An application in credit scoring. *European Journal of Operational Research* 261: 656–65. [CrossRef]

Malibari, Nadeem, Iyad Katib, and Rashid Mehmood. 2023. Systematic Review on Reinforcement Learning in the Field of Fintech. Available online: https://arxiv.org/pdf/2305.07466 (accessed on 31 October 2024).

Meltzer, Rachel. 2023. What Is Random Forest? Available online: https://careerfoundry.com/en/blog/data-analytics/what-is-random-forest/ (accessed on 31 October 2024).

Naik, K. S. 2021. Predicting Credit Risk for Unsecured Lending: A Machine Learning Approach. *arXiv* arXiv:2110.02206.

Nazarenko, E., V. Varkentin, and T. Polyakova. 2019. Features of Application of Machine Learning Methods for Classification of Network Traffic (Features, Advantages, Disadvantages). Paper presented at the 2019 International Multi-Conference on Industrial Engineering and Modern Technologies (FarEastCon), Vladivostok, Russia, October 1–4.

Sajumon, Akshatha. 2015. 5 Types of Credit Card Users You Definitely Fall Under. Available online: https://blog.bankbazaar.com/5-types-of-credit-card-users-you-definitely-fall-under/ (accessed on 31 October 2024).

Sariannidis, Nikolaos, Stelios Papadakis, Alexandros Garefalakis, Christos Lemonakis, and Tsioptsia Kyriaki-Argyro. 2020. Default avoidance on credit card portfolios using accounting, demographical and exploratory factors: Decision making based on machine learning (ML) techniques. *Annals of Operations Research* 294: 715–39. [CrossRef]

Sayjadah, Yashna, Ibrahim Abaker Targio Hashem, Faiz Alotaib, and Khairl Azhar Kasmiran. 2018. Credit Card Default Prediction using Machine Learning Techniques. Paper presented at the 2018 Fourth International Conference on Advances in Computing, Communication & Automation (ICACCA), Subang Jaya, Malaysia, October 26–28.

Sharma, Pranshu. 2023. Beginner's Guide To Decision Tree Classification Using Python. Available online: https://www.analyticsvidhya.com/blog/2021/04/beginners-guide-to-decision-tree-classification-using-python/ (accessed on 31 October 2024).

Shi, Si, Rita Tse, Wuman Luo, Stefano D'Addona, and Giovanni Pau. 2022. Machine learning-driven credit risk: A systemic review. *Neural Computing and Applications* 34: 14327–39. [CrossRef]

Sumiea, Ebrahim Hamid, Said Jadid Abdulkadir, Hitham Seddig Alhussian, Safwan Mahmood Al-Selwi, Alawi Alqushaibi, Mohammed Gamal Ragab, and Suliman Mohamed Fati. 2024. Deep deterministic policy gradient algorithm: A systematic review. *Heliyon* 10: e30697. [CrossRef] [PubMed]

Sun, Ting, and Miklos A. Vasarhelyi. 2018. Predicting credit card delinquencies: An application of deep neural networks. *Intelligent Systems in Accounting, Finance and Management* 25: 174–89. [CrossRef]

Tian, Zhenya, Jialiang Xiao, Haonan Feng, and Yutian Wei. 2020. Credit Risk Assessment based on Gradient Boosting Decision Tree. *Procedia Computer Science* 174: 150–60. [CrossRef]

Ullah, Mohammad Aman, Mohammad Manjur Alam, Shamima Sultana, and Rehana Sultana Toma. 2018. Predicting Default Payment of Credit Card Users: Applying Data Mining Techniques. Paper presented at the 2018 International Conference on Innovations in Science, Engineering and Technology (ICISET), Chittagong, Bangladesh, October 27–28.

Wang, Kui, Meixuan Li, Jingyi Cheng, Xiaomeng Zhou, and Gang Li. 2022. Research on personal credit risk evaluation based on XGBoost. *Procedia Computer Science* 199: 1128–35. [CrossRef]

Xia, Yufei, Chuanzhe Liu, Bowen Da, and Fangming Xie. 2018. A novel heterogeneous ensemble credit scoring model based on bstacking approach. *Expert Systems with Applications* 93: 182–99. [CrossRef]

Xu, Che, and Shuwen Zhang. 2024. A Genetic Algorithm-based sequential instance selection framework for ensemble learning. *Expert Systems with Applications* 236: 121269. [CrossRef]

Zhu, Lin, Dafeng Qiu, Daji Ergu, Cai Ying, and Kuiyi Liu. 2019. A study on predicting loan default based on the random forest algorithm. *Procedia Computer Science* 162: 503–13. [CrossRef]

risks

MDPI

Article

Evaluating Volatility Using an ANFIS Model for Financial Time Series Prediction

Johanna M. Orozco-Castañeda [1,*,†], Sebastián Alzate-Vargas [2,†] and Danilo Bedoya-Valencia [3]

[1] Instituto de Matemáticas, Universidad de Antioquia, Calle 67 No. 53-108, Medellín 050010, Colombia
[2] Departamento de Ciencias Matemáticas, Universidad de Puerto Rico Recinto Mayagüez, Mayagüez P.O. Box 9000, Puerto Rico; sebastian.alzate@upr.edu
[3] Independent Researcher, Medellín 050021, Colombia; dabedoyava@unal.edu.co
* Correspondence: johanna.orozco@udea.edu.co
† These authors contributed equally to this work.

Abstract: This paper develops and implements an Autoregressive Integrated Moving Average model with an Adaptive Neuro-Fuzzy Inference System (ARIMA-ANFIS) for BTCUSD price prediction and risk assessment. The goal of these forecasts is to identify patterns from past data and achieve an understanding of the future behavior of the price and its volatility. The proposed ARIMA-ANFIS model is compared with a benchmark ARIMA-GARCH model. To evaluated the adequacy of the models in terms of risk assessment, we compare the confidence intervals of the price and accuracy measures for the testing sample. Additionally, we implement the diebold and Mariano test to compare the accuracy of the two volatility forecasts. The results revealed that each volatility model focuses on different aspects of the data dynamics. The ANFIS model, while effective in certain scenarios, may expose one to unexpected risks due to its underestimation of volatility during turbulent periods. On the other hand, the GARCH(1,1) model, by producing higher volatility estimates, may lead to excessive caution, potentially reducing returns.

Keywords: optimization; dynamic systems; data modeling; forecasting; time series; fuzzy systems; soft computing; adaptive systems

MSC: 37M10; 37N40; 62A86

1. Introduction

Forecasting time series presents a significant challenge due to the inherent complexity and unpredictability of dynamic data. The primary difficulty lies in the fact that these data are generated via an unknown process, often perceived as random. Despite this, researchers strive to approximate this elusive data generation mechanism by analyzing observed data and patterns and applying a variety of models (Hyndman and Athanasopoulos 2018).

In time series analysis, there are two main features to model: the conditional mean and the conditional variance, also known as volatility. A model for the conditional mean allows us to explain and predict the behavior of the time series, while a model for the volatility enables the evaluation of the risk associated with the mean prediction, providing a confidence interval. This ability to forecast volatility is crucial for decision-making (Hamilton 2020).

The concept of time series in a probabilistic framework was first introduced by the Scottish statistician George Yule at the beginning of the 20th century. He defined a time series as a realization of a stochastic process, which is a set of random variables $Y = \{Y_t : t \in T\}$, where Y_t is a random variable and T is the index set. Generally, the index t is interpreted as time, and Y_t represents the state of the process at time t, with t typically being an integer.

In the formal approach to modeling a time series, a vector of lagged variables $X_{t-1} = (Y_{t-1}, Y_{t-2}, \cdots, Y_{t-p})^T$, $t > p$, is used from the process $\{Y_t\}$, and a regression given by Equation (1) is employed. That is,

$$Y_t = f(X_{t-1}, \theta) + \varepsilon_t, \tag{1}$$

where ε_t is a white noise process. This regression defines a model for the conditional mean and variance of the process given the available information up to time $t-1$. A general specification for a join model is as follows:

$$\mathrm{E}(Y_t|x_{t-1}) = f(X_{t-1}, \theta) \tag{2}$$

$$\mathrm{Var}(Y_t|x_{t-1}) = \mathrm{Var}(\varepsilon_t) = \sigma_t^2. \tag{3}$$

Here, θ is a vector of parameters that must be estimated based on certain criteria. Different functional forms of f give rise to various models for the process under study; additionally, different models can be used to represent the conditional variance.

The Autoregressive Integrated Moving Average (ARIMA) models are the most commonly used to model the conditional mean, also known as Box–Jenkins methodology (Hyndman and Athanasopoulos 2018; Wei 2006). These models use variations and regressions in the data to find patterns and make predictions for future values. The Generalized Autoregressive Conditional Heteroskedasticity (GARCH) model is a statistical model that is mostly used to analyze and forecast the volatility of time series data, particularly in financial markets; see for instance Bollerslev (2023). It extends the Autoregressive Conditional Heteroskedasticity (ARCH) model by incorporating lagged values of both the variance and the squared residuals, providing a more flexible and comprehensive framework for modeling the volatility clustering observed in financial time series (Engle and Patton 2007; Poon and Granger 2003).

Alternative models are often proposed to predict the conditional mean, conditional variance, or both measures jointly. For instance, to model the volatility, attractive options include the Takagi–Sugeno (T-S) fuzzy system, also known as the Sugeno fuzzy model, a type of fuzzy inference system developed by Takagi and Sugeno in 1985 (Takagi and Sugeno 1985) that has been applied to time series analysis and forecasting. This model is widely used in various applications due to its effectiveness in handling nonlinear systems and its ability to produce smooth control actions (Alenezy et al. 2023; Tsai et al. 2019).

Recent research with applications in financial data modeling have explored joint models for both conditional mean and variance. These include hybrid models that combine traditional statistical methods with machine learning and neural networks. Goodell et al. (2021) investigated the application of artificial intelligence and machine learning in financial time series analysis, identifying main areas such as portfolio optimization and investment, fraud and financial distress detection, forecasting, and financial planning. These authors highlight the transformative impact of these technologies and suggest some future research directions. Wang et al. (2013) proposed a hybrid ARIMA-GARCH model with a neural network component to capture non-linear relationships and improve forecast accuracy. These hybrid models leverage the strengths of both traditional and modern approaches, providing more robust and accurate predictions for financial time series.

In this work, we propose an Autoregressive Integrated Moving Average model with an Adaptive Neuro-Fuzzy Inference System (ARIMA-ANFIS) for stock price predictions and risk assessment. We also estimate an ARIMA-GARCH model as a benchmark to compare the performance of the proposed model. The Diebold–Mariano test, as proposed by Diebold and Mariano (1995), is applied to evaluate the predictive accuracy between the two models. The ARIMA models have a vast range of applications, and their properties are well known. Furthermore, several studies have demonstrated that ANFIS models can be successfully used for time series modeling due to their high flexibility and ability to model nonlinear dynamics. The ANFIS models are effective in modeling and predicting complex variables in various applications such as stock index, CO_2 emissions, global temperature,

the COVID-19 pandemic, and customer satisfaction; for more details, see Huarng and Yu (2005); Jiang et al. (2024); Jithendra and Sharief Basha (2023); Khan and Khan (2019). Although the specification process of the ANFIS model is not fully available in an analytical way, assessing different scenarios of this model with empirical applications allows us to make meaningful inferences.

The motivation for developing and implementing a joint ARIMA-ANFIS model is to forecast the daily closing price of BTC/USD and assess its associated risk in the stock market. The primary goal of these forecasts is to identify patterns from past data and gain an understanding of the future behavior of the price and its volatility. It is clear that to make profits in the context of trading, it is necessary to evaluate the risk of an asset and buy/sell an asset at a given price and close the trade at a higher/lower price. Therefore, having reliable forecasts that increase the probability of predicting price movements is key to maximizing profits. While forecasting stock prices is challenging due to the numerous factors influencing market behavior, it remains possible to identify patterns that provide valuable insights into future trends.

The objectives of this work include: understanding and analyzing ARIMA family models and ANFIS, carrying out the specification process for the ARIMA-ANFIS and ARIMA-GARCH models step by step, and applying them to real financial time series. We describe the ARIMA, GARCH, and ANFIS models, then we propose a join estimation process for the ARIMA-ANFIS and ARIMA-GARCH models, applied to real time series, followed by a comparative analysis of both approaches.

This paper is divided into five sections. Section 2 discusses several models used in this article and their properties. Section 3 outlines the steps for model formulation. The application and forecasting results are presented in Section 4. Finally, Section 5 concludes this paper.

2. Preliminaries

In this section, we provide a concise overview of the ARIMA, GARCH, and ANFIS models, along with an introduction to fuzzy logic, which forms the foundation for the ANFIS framework.

2.1. ARIMA Models

The ARIMA models are a class of statistical models, proposed by George Box and Gwilym Jenkins in the 1970s, that revolutionized the analysis and forecasting of time series by focusing on modeling the conditional mean of a time series. Known as the Box–Jenkins methodology, these models belong to the ARIMA family and are renowned for their ability to provide accurate forecasts in univariate time series. The Box–Jenkins methodology consists of four iterative steps: identification, parameter estimation, diagnostic checking, and forecasting. In the identification phase, it is of utmost importance to transform the original series into a stationary series, as stationarity is the fundamental condition for effectively constructing an ARIMA model (Hyndman and Athanasopoulos 2018).

The ARIMA models are recognized for their robustness and efficiency in forecasting financial time series, especially for short-term predictions, often outperforming the most popular artificial neural network techniques (Khashei and Bijari 2011). These models have been widely used in economics and finance. Various studies have employed ARIMA models for forecasting, as mentioned previously. ARIMA modeling is essentially an exploratory, data-oriented approach that allows for fitting an appropriate model based on the structure of the data. By using autocorrelation and partial autocorrelation functions, it is possible to model the stochastic nature of the time series. This facilitates the identification of trends, random variations, periodic components, cyclic patterns, and serial correlations. As a result, forecasts of future values of the series can be obtained with a certain degree of accuracy.

The ARIMA(p, d, q) model can be written as follows:

$$\nabla^d \dot{Y}_t = \alpha + \varphi_1 \nabla^d \dot{Y}_{t-1} + \varphi_2 \nabla^d \dot{Y}_{t-2} + \ldots + \varphi_p \nabla^d \dot{Y}_{t-p} + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \ldots + \theta_q \varepsilon_{t-q} + \varepsilon_t,$$

which can also be expressed as follows:

$$\Phi_p(B)(1-B)^d \dot{Y}_t = \Theta_q(B)\varepsilon_t,$$

where $\nabla^d = (1-B)^d$, p is the number of lags in the model, q is the size of the moving average window, and d corresponds to the unit roots of the process; the number of times the time series must be differenced to achieve stationarity. The model parameters $\varphi_1, \ldots, \varphi_p$, $\theta_1, \ldots, \theta_q$ must satisfy certain conditions to ensure the stationarity and invertibility of the process. α is a constant, and $\{\varepsilon_t\}$ represents a sequence of identically distributed independent random variables with zero mean and constant variance, commonly referred to as a white noise process. For example, ARIMA(0,0,0) represents white noise, and ARIMA(0,1,0) with/without a constant corresponds to a random walk with/without drift.

The selection of the p, d, and q values is part of the model identification process, which is conducted using statistical criteria, the autocorrelation and partial autocorrelation functions, and statistical tests such as the Dickey–Fuller test for unit roots (Dickey and Fuller 1979).

2.2. GARCH Models

One of the most used models for statistical modeling and forecasting the conditional volatility is the generalized autoregressive conditional heteroskedastic (GARCH) approach of Bollerslev (2023). In a GARCH(p, q) model, the current variance, σ_t^2, is a function of the past squared shocks, $\{a_{t-i}^2; i = 1, \ldots, p\}$, and the past variances, $\{\sigma_{t-j}^2; j = 1, \ldots, q\}$:

$$\sigma_t^2 = \kappa + \sum_{i=1}^{p} \alpha_i a_{t-i}^2 + \sum_{j=1}^{q} \beta_j \sigma_{t-j}^2$$

with

$$a_t = \sigma_t \varepsilon_t \text{ and } \varepsilon_t \sim N(0,1)$$

where the parameters κ, α_i, and β_j are subject to the following restrictions: $\kappa > 0$, $\alpha_i \geq 0$, $\beta_j \geq 0$, and $\sum_{i=1}^{\max(p,q)} (\alpha_i + \beta_i) < 1$.

2.3. Fuzzy Logic

Fuzzy logic, introduced by Zadeh in 1965 (Zadeh 1965), provides a mathematical framework for representing and managing uncertainty and vagueness. It offers formal tools such as fuzzy sets, if–then rules, and fuzzy arithmetic. In fuzzy logic, the degree of fuzzy membership signifies the similarity between events, where the exact properties of these events are not precisely defined. Membership degrees in fuzzy logic range continuously between 0 (completely false) and 1 (completely true). Zadeh's pioneering work on fuzzy sets laid the foundation of fuzzy set theory. The core concept is that an element can belong to a set with a degree of membership, making propositions not strictly true or false but instead partially true or false. Following fuzzification of values, linguistic rules are applied to derive outputs, which may either retain their fuzzy nature or be defuzzified to yield a discrete (numerical) value.

Working with fuzzy sets begins with defining membership function and establishing the domain of the problem, which involves selecting appropriate functions to represent it.

Linguistic rules are employed to connect inputs with outputs. A fuzzy rule is symbolically represented as follows:

IF <fuzzy statement> THEN <fuzzy statement>

where <fuzzy statement> could be an expression in natural language.

In fuzzy systems featuring fuzzy premises, all rules are partially activated, and the consequent is true to a certain extent. The process used to compute an output value for a given input is known as fuzzy inference. This process distinguishes between two primary types: the Mamdani model, proposed by Mamdani and Assilian (Mamdani and Assilian 1999) in 1975, and the TSK model (Takagi, Sugeno, and Kang), introduced by Takagi and Sugeno (1985) as an alternative to the Mamdani model. In the present study, the TSK model is employed.

In Takagi–Sugeno fuzzy systems, the rules are structured as follows:

$$\text{IF } x \in A \text{ and } y \in B \text{ THEN } z = f(x, y),$$

where x and y represent the input variables, and A, B are fuzzy sets associated with membership functions. These membership functions can take various functional forms chosen from a wide range of options.

2.4. Adaptive Neuro Fuzzy Inference Systems—ANFIS

Adaptive Neuro-Fuzzy Inference Systems (ANFISs) are computational techniques from the domain of soft computing. Soft computing techniques are designed to model and handle uncertainty, approximation, and imprecision in problem-solving. Unlike traditional (hard) computing, which relies on binary logic and crisp values, soft computing incorporates methodologies that allow for flexibility and tolerance for imprecision, making it suitable for complex, real-world problems where exact solutions are difficult or impossible to find.

ANFISs are recognized for their adaptability and effectiveness in modeling complex relationships, especially in nonlinear systems and time series prediction (Walia et al. 2015). These models have found wide applications in various fields due to their ability to accurately capture intricate patterns in data. Similar to ARIMA models, ANFIS modeling involves an exploratory approach that utilizes data-driven techniques to tailor the model to the specific characteristics of the dataset. By integrating fuzzy logic principles with neural network architectures, ANFIS models can effectively handle the uncertainties and nonlinearity present in time series data (Talebizadeh and Moridnejad 2011). This capability allows ANFIS models to uncover trends, periodic components, and other complex patterns that influence the behavior of financial markets.

The Takagi–Sugeno-type fuzzy systems from Takagi and Sugeno (1985) have been used in modeling time series and predicting the mean and volatility (Sahiner et al. 2023; Venugopal et al. 2024). These models have demonstrated the ability to provide accurate forecasts and address the challenges inherent in volatility prediction. ANFIS is a type of Takagi–Sugeno fuzzy system that combines the fuzzy logic principles of Takagi–Sugeno models with neural network structures. These model combine the interpretability of fuzzy systems with the learning capabilities of neural networks.

The fundamental characteristic of ANFIS models is their ability to partition each input variable into two or more regions. The structure of ANFIS models is illustrated in Figure 1, where X and Y are the independent variables, and z represents the defuzzification. The domain of X is divided into regions A_1 and A_2, and the domain of Y is divided into regions B_1 and B_2. Consequently, the domain of the system, the xy plane, is divided into regions $1, 2, 3$, and 4. Each of these regions have been assigned a linear model of the form $z = ax + by + c$ as the consequent function.

Finally, the model is represented by a set of rules, where the antecedent determines the region that the point to be evaluated belongs to, while the consequent corresponds to the linear model. One of the advantages of ANFIS models is that a point can simultaneously belong to two or more regions. Consequently, the value of z is calculated considering the linear models of each region. As an example, this area of ambiguity is shown in gray in Figure 1. In this case, the membership of a point to a region is determined through a fuzzy set, which has its membership function, $\mu_i(x)$, indicating the degree to which it is associated with the fuzzy set S_i, as mentioned in the previous section. The training process of an ANFIS involves the estimation of premise and consequence parameters using an optimization algorithm. The choice of optimization method is crucial for achieving optimal results, as highlighted by Karaboga and Kaya (2019).

For this example, the inference process performed to calculate the value of z given an input is as follows:

- Calculate the membership functions $\mu_1, \mu_2, \mu_3, \mu_4$ for each point in the xy plane.

- For each rule, estimate w_j, which is defined as the firing strength of each rule. It can be calculated as a weight of the membership values or by multiplying the membership function values.
- Establish the percentage contribution of each rule to the final solution: $\bar{w}_j = \frac{w_j}{\sum_{i=1}^{4} w_i}$.
- Finally, calculate the system's output as $\hat{f}_t = \sum_i (w_i f_i)$, where f_i is typically a linear combination of the variables in the consequent.

The volatility models play a fundamental role in financial decision-making by providing insights into the uncertainty and potential future movements of asset prices. They must be able to quantify and forecast the variability associated with the returns of a financial time series. Despite being a topic of interest for many researchers (Poon and Granger 2003), there are several challenges in obtaining accurate volatility forecasts. These challenges include the fact that volatility is a non-observable feature, its estimator has a changing variance over time, and it exhibits clusters of similar variances, heavy-tailed distributions, and non-linear and non-stationary behavior, among other complexities (Bollerslev and Engle 1993; Poon and Granger 2003). The ability of ANFIS to model complex behaviors and nonlinear systems, especially for predicting volatility, is indeed a promising tool. ANFIS provides a sophisticated approach for predicting volatility by leveraging the strengths of neural networks and fuzzy logic.

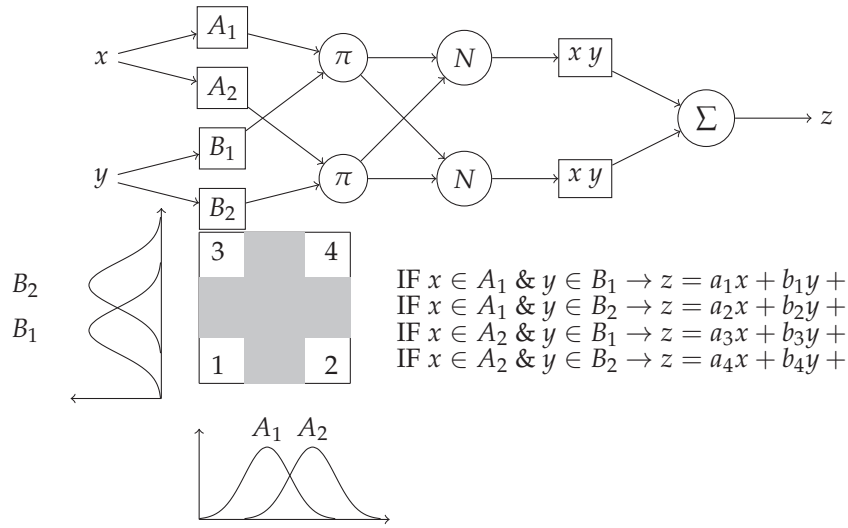


Figure 1. Typical ANFIS structure. Adapted from Jang (1993).

3. Specification of ARIMA-ANFIS and ARIMA-GARCH Models

The ARIMA-ANFIS and ARIMA-GARCH models are estimated to predict the time series and evaluate the quality of this prediction by computing the confidence intervals using the volatility estimated from the ANFIS and GARCH models.

First, we model the conditional mean using the ARIMA model and obtain the residuals. Then, we check the correlation in the squared residuals to determine the appropriateness of applying the ANFIS and GARCH model for capturing the conditional variance of the time series.

3.1. Identification Process for the Conditional Mean Model

We perform an automatic model identification process to determine which model best explains the BTC/USD price. For the selected model, we perform validation tests for the residuals, a normality test, and finally, we determine the independence of the residual using the Ljung–Box test.

The forecast values of the time series are computed using the library `forecast` (from R package) to generate predictions from the fitted ARIMA model.

The framework for the training and testing process is described as follows: initially, the identification process and the model estimation are made using the training sample,

which comprises 80% of the total dataset. Subsequently, the model is tested using the remaining 20% of the data, which is called the testing sample, by making predictions for 1 day ahead. This testing strategy employs an expanding window strategy, where the model is retrained at each step following the cross-validation (CV) guidelines for time series, as outlined by Bergmeir et al. (2018). For further references, please see Bergmeir et al. (2018).

3.2. Volatility with ANFIS Model

Let $\{\varepsilon_t\}$ be the time series of the residuals from the ARIMA model in the training sample. The ANFIS model consist of L fuzzy rules, written as follows:

$$\text{IF } x_t \in S_i \text{ THEN } f_i = f(\theta_i, x_t).$$

Here, $x_t = [\varepsilon_{t-1}^2, \varepsilon_{t-2}^2]$, with ε_{t-1} and ε_{t-2} being the lagged values of ε_t; S_i is a fuzzy set in the input space; f_i is the forecasted value for ε_t^2 using the i-th rule; and $f(\cdot,\cdot)$ is a function that takes the following form:

$$f_i(x_t) = a_i\varepsilon_{t-1}^2 + b_i\varepsilon_{t-2}^2 + c_i,\ i = 1, \ldots, L, \tag{4}$$

where a_i, b_i and c_i are parameters to be determined.

Note that each region into which the domain is divided is assigned a model of the form (4). It is necessary to define the membership function of x_t to the set S_i. In ANFIS, each set S_i is represented by its center m_i, and the value of the membership function, $\mu_i(x_t)$, is defined as a function of the distances from the point x_t to the center of the clusters or fuzzy sets.

As an example, Figure 2 presents the architecture of a model for an ANFIS with four fuzzy rules. For a 3D representation, refer to Aznarte and Benítez (2010), which illustrates the division of the plane into four fuzzy sets, along with their corresponding membership functions and the associated planes for each set. For a point in the plane, we obtain membership values for all the fuzzy sets and the defuzzified values given in Equation (4).

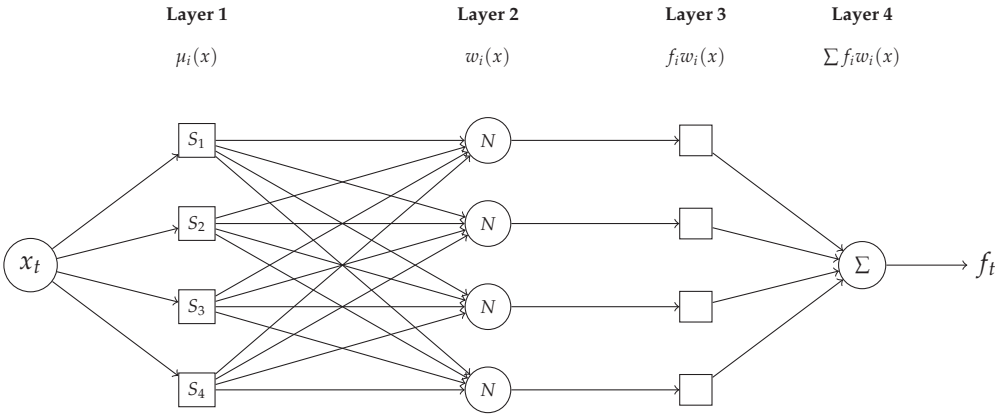


Figure 2. Architecture for an ANFIS with four rules.

This ANFIS model consist of four layers: fuzzification, rule evaluation, aggregation, and defuzzification. The optimization process typically employs a learning algorithm based on gradient descent methods. During training, the parameters of the membership functions and the network weights are adjusted to minimize the error between the actual and predicted values. The process in each layer is described below.

Layer 1—Fuzzification

In this layer, we compute the membership function μ_i from the point x_t to the fuzzy set S_i:

$$\mu_i = \mu_{S_i}(x_t) = \exp\left[-\frac{D_i^2}{d_i^2}\right] = \exp\left[-\frac{(\varepsilon_{t-1}^2 - m_{i1})^2 + (\varepsilon_{t-2}^2 - m_{i2})^2}{d_i^2}\right], \; i = 1, \ldots, L.$$

Here, d_i is the standard deviation of fuzzy set S_i, and D_i^2 is the squared distance from the point x_t to the center $m_i = [m_{i1}, m_{i2}]$. In other words,

$$D_i^2 = ||x_t - m_i||^2,$$

where $||\cdot||$ represents the norm. The membership function μ_i is equal to one when $x_t = m_i$ and decreases further as the distance from the point x_t to the center m_i increases. This reduction occurs when the fuzzy set S_i has very dispersed points or the current point x_t belongs to another fuzzy set.

Layer 2—Rule Evaluation

In this layer, each fuzzy rule's firing strength is determined based on the degree to which the point x_t matches the premises (the IF portion) of the rules. The firing strength w_i, which represents the degree of contribution of each rule to the final output, is given as follows:

$$w_i = \frac{\mu_i}{\sum_{j=1}^{L} \mu_j}, \; i = 1, \ldots, L.$$

Layer 3—Aggregation

The firing strengths from all rules are combined to produce a single aggregated output $w_i f_i$, for $i = 1, \ldots, L$.

Layer 4—Defuzzification

We compute the final output by combining the outputs of all the rules into a single crisp value:

$$f_t = \sum_{j=1}^{L} w_j f_j.$$

4. Application to a Real Time Series

In this section, we apply the ARIMA-ANFIS model to the daily price series of the BTCUSD currency pair (Bitcoin to US Dollar) using closing prices from 3 August 2023 to 31 July 2024. The time series consists of 364 daily data points. The ARIMA model is used to study the currency price series, and the residuals from this model, which exhibit non-constant variance, are modeled using the ANFIS and GARCH(1,1) model. Additionally, all results from this application were obtained using RStudio program.

4.1. Identification Process for the Conditional Mean Model

We conducted an identification and validation process for the ARIMA model using a training sample consisting of $n = 291$ observations. For model selection, we generated a variety of ARIMA(p,d,q) models, exploring combinations for $p, q = 0, 1, 2, 3$ and $d = 0, 1, 2$. We assessed their performance using multiple criteria, including the Akaike Information Criterion (AIC), log-likelihood, and the Mean Absolute Percentage Error (MAPE). Additionally, we conducted residual diagnostics to ensure the adequacy of the selected model. Through this evaluation process, we determined that the ARIMA(2,1,2) model provided reliability for our specific dataset. Using the selected model and the forecast package, we obtained fitted values for the training sample and predictions for the testing sample. Figure 3 displays the actual time series (black line), the fitted values (blue line), the predictions (green line), and a red dashed vertical line marking the separation between the training and testing samples.

To evaluate the performance of the ARIMA(2,1,2) model, we use the Root Mean Squared Error (RMSE), the Mean Absolute Error (MAE), and the Mean Absolute Percentage Error (MAPE) for both the training sample and the testing sample. The measures are calculated as follows:

$$\text{RMSE}_{\text{trainingsample}} = \sqrt{\frac{1}{n}\sum_{t=1}^{n} r_t^2}$$

$$\text{RMSE}_{\text{testingsample}} = \sqrt{\frac{1}{N-n}\sum_{t=n+1}^{N} e_t^2}$$

$$\text{MAE}_{\text{trainingsample}} = \frac{1}{n}\sum_{t=1}^{n} |r_t|$$

$$\text{MAE}_{\text{testingsample}} = \frac{1}{N-n}\sum_{t=n+1}^{N} |e_t|$$

$$\text{MAPE}_{\text{trainingsample}} = \frac{1}{n}\sum_{t=1}^{n} \left|\frac{r_t}{y_t}\right| \times 100$$

$$\text{MAPE}_{\text{testingsample}} = \frac{1}{N-n}\sum_{t=n+1}^{N} \left|\frac{e_t}{y_t}\right| \times 100,$$

where n is the training sample size, N is the length of the time series of the BTCUSD price, y_t is the actual BTCUSD price, r_t is the residual from ARIMA(2,1,2), and e_t is the forecast error at time t.

Table 1 presents the performance metrics of the ARIMA(2,1,2) model, evaluated based on RMSE, MAE, MAPE, and R^2. These metrics provide insights into the accuracy and reliability of the model. The RMSE for the testing sample is 1408.954 versus 1343.355 for the training sample, and the MAE is 1061.159 for the testing sample versus 869.390 for the training sample. The RMSE and MAE values are slightly lower for the training sample, but it is expected that the models will perform better on the data they were trained on. On the other side, the MAPE is lower on the testing set compared to the training set, indicating that the model has a good generalization capability. The MAPE values are less than 2%, which means that the predictions are, on average, 2% off from the actual values. This would be considered highly accurate. Additionally, although the R^2 value for the training sample gives extremes results, it still performs reasonably well on the testing data, with an R^2 value equal to 0.8856. The difference between the training and testing values could indicate some overfitting, but it is not excessive, given that the testing R^2 value is still high. The observed values in the accuracy measures suggest that the model is not overfitting and seems to generalize well from the training data to the testing data.

Figure 3. BTC/USD price vs forecast from ARIMA model.

Table 1. Accuracy measures for training and testing samples of ARIMA(2,1,2).

Sample	RMSE	MAE	MAPE	R^2
Training	1343.355	869.390	1.7998	0.9922
Testing	1408.954	1061.159	1.6486	0.8856

4.2. ANFIS and GARCH Estimation

Consider the time series $\{r_t\}$ formed by the residuals from the ARIMA(2,1,2) model of length $n = 291$ and the forecast errors produce by the ARIMA forecasts of length $N - n = 73$.

For the ANFIS estimation, we define four bivariate fuzzy sets, each represented by a corresponding fuzzy rule. The parameters to be estimated include the centers and standard deviations of the four fuzzy sets, as well as the parameters of the consequents corresponding to planes in 3D space. The membership function employed here is the Gaussian function, given as follows:

$$\mu_i = \exp\left[-\frac{||x_t - m_i||^2}{d_i^2}\right],$$

where d_i is the standard deviation of the i-th fuzzy set, and x_t is the point of residuals at time t, defined as $x_t = [\varepsilon_{t-1}^2, \varepsilon_{t-2}^2] = [r_{t-1}^2, r_{t-2}^2]$ and $m_i = [m_{i1}, m_{i2}]$.

We initialize the parameters of the ANFIS model with random uniform numbers. To optimize the objective function with the mean squared error, we employ the Nelder–Mead method using the `optim` function.

Figure 4 displays the scatterplot for the ordered pairs x_t. Note that there are large squared residuals at the first/second lag and small squared residuals at the second/first lag, as well as small values at both lags. The ANFIS algorithm divides the plane of the squared residuals in four fuzzy sets and performs the entire inference process based on the observations $\{r_t^2\}$ for $t = 1, \cdots, n$, which is called the training sample. It generates an estimate for the variance, $\hat{\sigma}_{n+1}^2$, which serves as a one-step-ahead forecast. This process is conducted iteratively, with each new observation r_t^2, for $t = n+1, \cdots, N-1$, being added sequentially. At each step, the ANFIS algorithm is re-estimated and applied to produce a one-step-ahead variance forecast for the subsequent time $t+1$ following an expanding window testing strategy.

For the GARCH(1,1) estimation, we use the library `rugarch` from the R program. The specification was `ugarchspec(variance.model = list(model = "sGARCH", garchOrder = c(1, 1)), mean.model = list(armaOrder = c(0, 0), include.mean = FALSE), distribution.model = "std")`.

The GARH(1,1) model is first estimated using the training sample $\{r_t^2\}$ $t = 1, \cdots, n$. Subsequently, a new observation is sequentially added to the time series, similar to the procedure describe above for the ANFIS process. We use the `ugarchforecast` function to generate one step ahead variance forecasts at each time $t+1$.

We use the standardized squared residuals from the ARIMA model as a proxy for the actual variance at each point. In Figure 5, these residuals are presented alongside the forecasts of ANFIS (blue line) and the GARCH(1,1) model (purple line); the red dashed vertical line marks the separation between the training and testing samples. Note that even when ANFIS is estimated by minimizing the MSE, the volatility predictions are lower than those from the GARCH model. Here, we observe that ANFIS is better at capturing periods of market stability, while the GARCH model excels in capturing high volatility. During calm periods, the GARCH model might overestimate risks, and during volatile periods, ANFIS might underestimate risks.

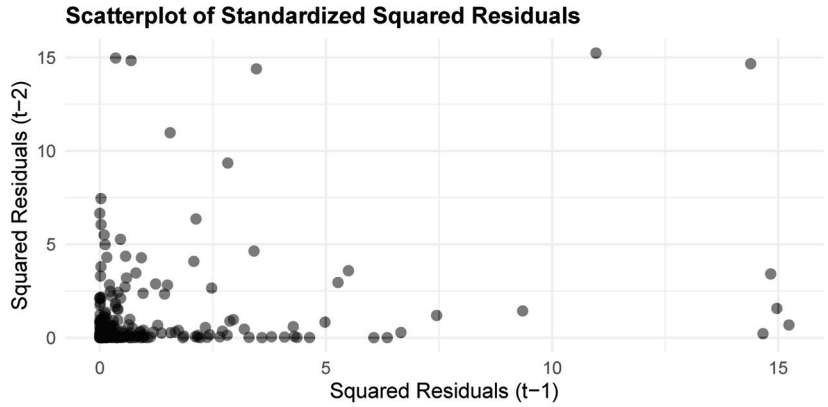


Figure 4. Scatterplot for squared returns.

Table 2 presents the MSE and MAE of the testing sample for ANFIS and GARCH(1,1). Even though the MSE for ANFIS is around 10% greater than the MSE for the GARCH(1,1) model, the MAE for ANFIS is around 16% smaller than the MAE for GARCH.

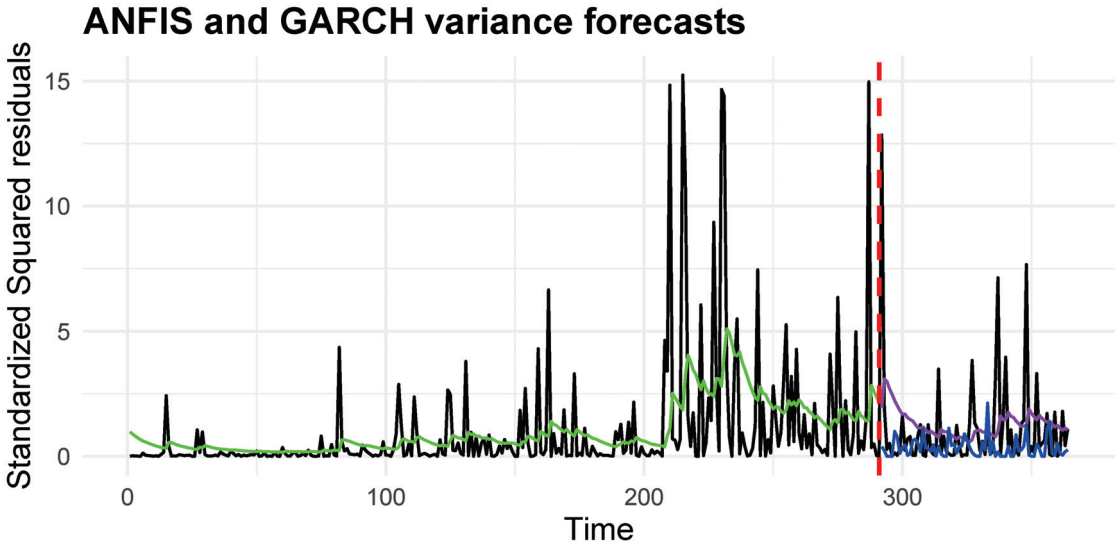


Figure 5. Predictions in the testing sample with ANFIS and GARCH.

Table 2. Accuracy measures for the testing sample of ANFIS and GARCH(1,1).

Model	MSE	MAE
ANFIS	4.5539	1.0780
GARCH(1,1)	4.2307	1.3172

Once we finished the ANFIS and GARCH(1,1) estimation processes, we obtained the 73 volatility forecasts for the residuals of the ARIMA model. These volatility forecasts, $\hat{\sigma}_t$, are used to compute the 95% confidence intervals for the price of the BTCUSD currency around $\hat{y}$ in the testing sample for both approaches as follows:

$$(\hat{y}_t - t_{0.975,N-n} \times \hat{\sigma}_t, \hat{y}_t + t_{0.975,N-n} \times \hat{\sigma}_t).$$

Here, $\hat{y}_i$, $i = n+1, \ldots, N$ are the price forecasts from the ARIMA model in the testing sample, and $t_{0.975,N-n}$ are the percentage points of the t distribution.

Figures 6 and 7 show the confidence intervals for the price predictions using the ANFIS and GARCH(1,1) models, respectively. The varying widths of the confidence

intervals are due to the standard deviation within each, which corresponds to the square root of the predicted variance and changes for each time t. It is also noteworthy that some prediction points generated by the ARIMA model have unusually narrow confidence intervals. However, a closer examination of the real data time series reveals that these points coincide with abrupt changes, as illustrated in Figure 6. This could be seen as a positive aspect for currency trading, as the ARIMA model would indicate high-risk operations at these points.

The results revealed that 61.64% of the time, the BTC/USD price fell within the predicted confidence intervals generated by the ANFIS model. In contrast, the GARCH(1,1) model captured the BTC/USD price within its predicted confidence intervals 94.52% of the time. This discrepancy can be attributed to the superior ability of the GARCH model to capture periods of high volatility, leading to larger estimates of volatility compared to ANFIS.

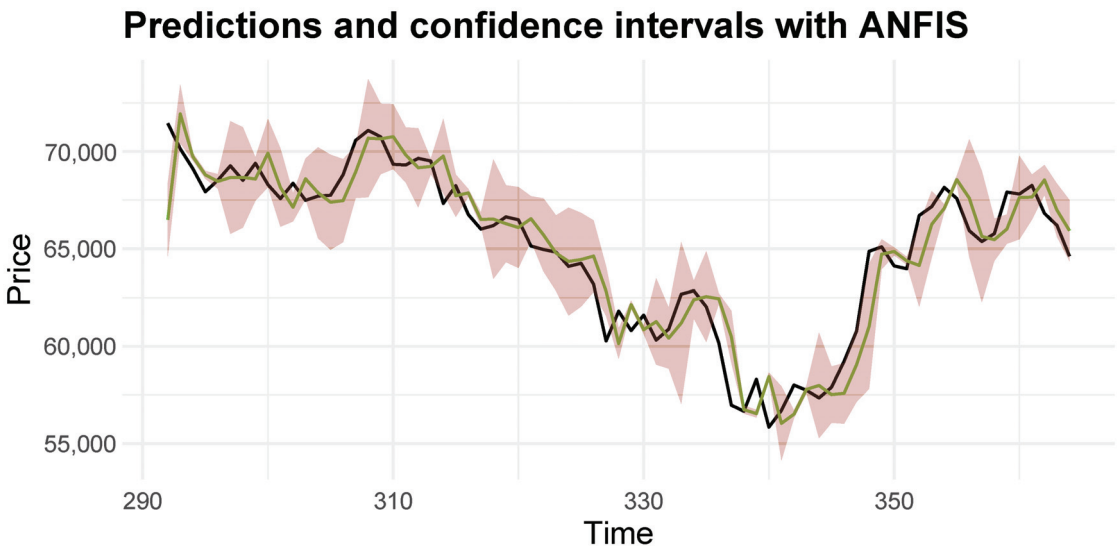


Figure 6. Predictions and confidence intervals in the testing sample with ANFIS.

We compared the predictive accuracy of the two competing volatility forecasts using the Diebold–Mariano test. Specifically, we evaluated the test based on two different loss functions: squared error loss and absolute error loss. The loss differential of the two competing models is calculated as follows:

$$d_t = L(e_{1,t}) - L(e_{2,t})$$

where $e_{1,t}$ and $e_{2,t}$ are the prediction errors from the ANFIS and GARCH models, respectively, and $L(\cdot)$ represent the loss function.

In a two-sided Diebold and Mariano test, the high p-value for the squared error loss function, as shown in Table 3, indicates no significant difference between the models in capturing large errors, suggesting that neither model is more accurate during high-volatility periods. However, the p-value for the absolute error loss function, suggests a significant difference when evaluating smaller errors. The mean of the loss differential indicates that ANFIS may better capture low-volatility conditions. Thus, the Diebold–Mariano test suggests that the models perform differently across volatilty regimes: ANFIS is more accurate in low-volatility environments, while neither model shows superiority in high-volatility conditions. This finding aligns with the previous analysis based on the accuracy measures presented in Table 2 and the computed confidence intervals.

Table 3. Predictive accuracy of ANFIS vs. GARCH model based on the Diebold and Mariano test.

Loss Function	*p*-Value
Absolute error loss	0.031 *
Squared error loss	0.590

* means a significant difference between the models.

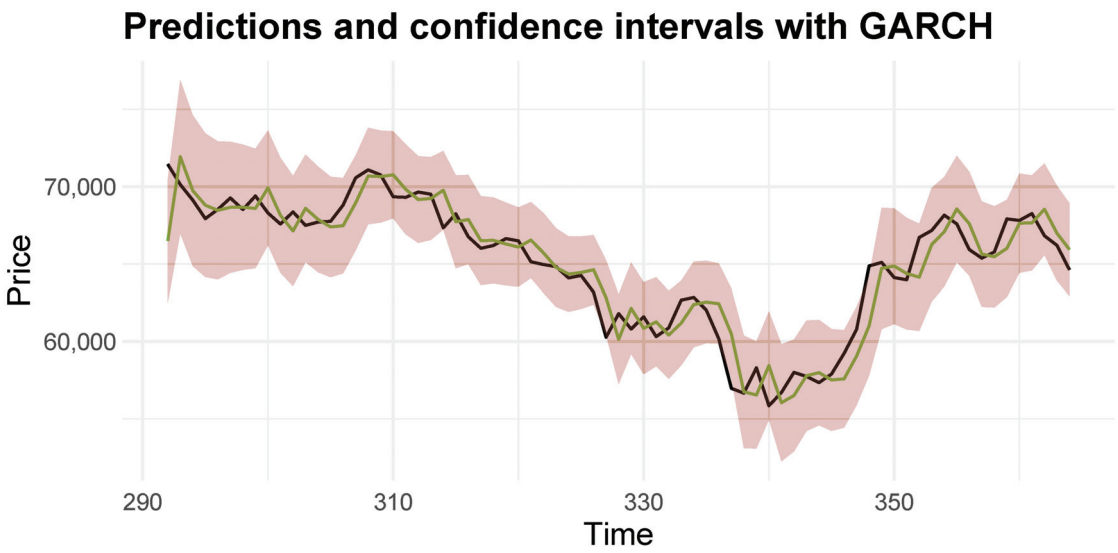


Figure 7. Predictions and confidence intervals for the testing sample using GARCH(1,1).

5. Conclusions

In this work, we addressed a complex and challenging problem in modeling financial time series. Through a novel and interesting approach, we provided insights and solutions that contribute significantly to the understanding of this issue. Our methodology, which incorporates the processes of identification, specification, estimation, and validation for the ARIMA model and adaptive neuro-fuzzy inference systems, demonstrates the potential for effective application in real-world scenarios.

We implemented the Adaptive Neuro-Fuzzy Inference System (ANFIS), a prominent technique within the domain of soft computing. ANFIS integrates the Takagi–Sugeno–Kang (TSK) fuzzy inference model with neural network methodologies, leveraging fuzzy set theory, IF–THEN fuzzy rules, and fuzzy reasoning. This hybrid system employs diagrams and a connectionist representation, inspired by the functioning of the brain, to effectively model complex and nonlinear systems.

In this article, the ARIMA model is employed to capture the conditional mean, while the ANFIS methodology is used to model the conditional variance of financial series, specifically the daily BTCUSD price. It is noteworthy that conditional variance in these time series is typically modeled using GARCH models; hence, applying ANFIS methodology in this context is innovative. By combining the econometric approach (ARIMA) with the soft computing technique (ANFIS), we jointly model both the conditional mean and conditional variance, creating a hybrid ARIMA-ANFIS model.

The comparison between the benchmark ARIMA-GARCH model and the proposed ARIMA-ANFIS model reveals that each model captures different aspects of data dynamics. While the ANFIS model is effective in certain scenarios, it may underestimate volatility during turbulent periods, potentially exposing users to unexpected risks, as illustrated between observations 335 and 340 in Figure 6. Conversely, the GARCH(1,1) model, by generating higher volatility estimates, might lead to excessive caution, potentially reducing returns. This highlights the trade-offs between the two models: ANFIS offers a more conservative approach in stable markets, whereas GARCH(1,1) provides a robust defense against high

volatility, but at the cost of possibly missing out on opportunities during calmer periods, as shown in Figure 7.

Parameter optimization for the ANFIS model proved to be a time-consuming procedure, highlighting the need for more efficient optimization techniques. As future work, the implementation of evolutionary algorithms could be explored. Evolutionary algorithms, with their robustness and global search capabilities, have the potential to significantly streamline the optimization process by efficiently navigating the complex search space of the ANFIS parameters.

The proposed ARIMA-ANFIS model adequately captured some of the dynamics in the treated financial time series, providing good forecasts and confidence intervals in most cases. Additionally, testing this model using other time series exhibiting non-constant conditional variance could be considered.

Author Contributions: Conceptualization, J.M.O.-C., S.A.-V. and D.B.-V.; methodology, J.M.O.-C., S.A.-V. and D.B.-V.; investigation, J.M.O.-C., S.A.-V. and D.B.-V.; writing—original draft preparation, J.M.O.-C., S.A.-V. and D.B.-V.; writing—review and editing, J.M.O.-C., S.A.-V. and D.B.-V. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Publicly available datasets were analyzed in this study. This data can be found at Yahoo Finance.

Acknowledgments: We thank Juan David Velásquez from Universidad Nacional de Colombia for their invaluable mentorship and support throughout the research journey. We would also like to thank the referees for taking the time and effort necessary to review this article. We sincerely appreciate all valuable comments and suggestions, which helped us to improve the quality of this article.

Conflicts of Interest: The authors declare no conflicts of interest.

References

Alenezy, Abdullah H., Mohd Tahir Ismail, Sadam Al Wadi, and Jamil J. Jaber. 2023. Predicting stock market volatility using modwt with hyfis and fs.hgd models. *Risks* 11: 121. [CrossRef]

Aznarte, José Luis, and José Manuel Benítez. 2010. Equivalences between neural-autoregressive time series models and fuzzy systems. *IEEE Transactions on Neural Networks* 21: 1434–44. [CrossRef]

Bergmeir, Christoph, Rob J. Hyndman, and Bonsoo Koo. 2018. A note on the validity of cross-validation for evaluating autoregressive time series prediction. *Computational Statistics & Data Analysis* 120: 70–83.

Bollerslev, Tim. 2023. Reprint of: Generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics* 234: 25–37. [CrossRef]

Bollerslev, Tim, and Robert F. Engle. 1993. Common Persistence in Conditional Variances. *Econometrica* 61: 167–86. [CrossRef]

Dickey, David A., and Wayne A. Fuller. 1979. Distribution of the estimators for autoregressive time series with a unit root. *Journal of the American Statistical Association* 74: 427–31.

Diebold, Francis X., and Roberto S. Mariano. 1995. Comparing predictive accuracy. *Journal of Business & Economic Statistics* 13: 253–63.

Engle, Robert F., and Andrew J. Patton. 2007. What good is a volatility model? In *Forecasting Volatility in the Financial Markets*. Amsterdam: Elsevier, pp. 47–63.

Goodell, John W., Satish Kumar, Weng Marc Lim, and Debidutta Pattnaik. 2021. Artificial intelligence and machine learning in finance: Identifying foundations, themes, and research clusters from bibliometric analysis. *Journal of Behavioral and Experimental Finance* 32: 100577. [CrossRef]

Hamilton, James D. 2020. *Time Series Analysis*. Princeton: Princeton University Press.

Huarng, Kunhuang, and Hui-Kuang Yu. 2005. A type 2 fuzzy time series model for stock index forecasting. *Physica A: Statistical Mechanics and its Applications* 353: 445–62. [CrossRef]

Hyndman, Rob J., and George Athanasopoulos. 2018. *Forecasting: Principles and Practice*. Melbourne: OTexts.

Jang, J.S. Roger. 1993. Anfis: Adaptive-network-based fuzzy inference system. *IEEE Transactions on Systems Man and Cybernetics* 23: 665–85. [CrossRef]

Jiang, Huimin, Farzad Sabetzadeh, and Chen Zhang. 2024. An intelligent adaptive neuro-fuzzy inference system for modeling time-series customer satisfaction in product design. *Systems* 12: 224. [CrossRef]

Jithendra, Thandra, and Shaik Sharief Basha. 2023. A hybridized machine learning approach for predicting COVID-19 using adaptive neuro-fuzzy inference system and reptile search algorithm. *Diagnostics* 13: 1641. [CrossRef]
Karaboga, Dervis, and Ebubekir Kaya. 2019. Adaptive network based fuzzy inference system (ANFIS) training approaches: A comprehensive survey. *Artificial Intelligence Review* 52: 2263–93. [CrossRef]
Khan, Muhammad Zahir, and Muhammad Farid Khan. 2019. Application of anfis, ann and fuzzy time series models to co2 emission from the energy sector and global temperature increase. *International Journal of Climate Change Strategies and Management* 11: 622–42. [CrossRef]
Khashei, Mehdi, and Mehdi Bijari. 2011. A novel hybridization of artificial neural networks and arima models for time series forecasting. *Applied Soft Computing* 11: 2664–75. [CrossRef]
Mamdani, Ebrahim H., and Sedrak Assilian. 1999. An experiment in linguistic synthesis with a fuzzy logic controller. *International Journal of Human-Computer Studies* 51: 135–47. [CrossRef]
Poon, Ser-Huang, and Clive W. J. Granger. 2003. Forecasting Volatility in Financial Markets: A Review. *Journal of Economic Literature* 41: 478–539. [CrossRef]
Sahiner, Mehmet, David G. McMillan, and Dimos Kambouroudis. 2023. Do artificial neural networks provide improved volatility forecasts: Evidence from Asian markets. *Journal of Economics and Finance* 47: 723–62. [CrossRef]
Takagi, Tomohiro, and Michio Sugeno. 1985. Fuzzy identification of systems and its applications to modeling and control. *IEEE Transactions on Systems, Man, and Cybernetics* SMC-15: 116–32. [CrossRef]
Talebizadeh, Mansour, and Ali Moridnejad. 2011. Uncertainty analysis for the forecast of lake level fluctuations using ensembles of ann and anfis models. *Expert Systems with Applications* 38: 4126–35. [CrossRef]
Tsai, Ming-Chi, Ching-Hsue Cheng, and Meei-Ing Tsai. 2019. A multifactor fuzzy time-series fitting model for forecasting the stock index. *Symmetry* 11: 1474. [CrossRef]
Venugopal, Ravi, Chinnadurai Veeramani, and S. Muruganandan. 2024. An effective approach for predicting daily stock trading decisions using fuzzy inference systems. *Soft Computing* 28: 3301–19. [CrossRef]
Walia, Navneet, Harsukhpreet Singh, and Anurag Sharma. 2015. Anfis: Adaptive neuro-fuzzy inference system-a survey. *International Journal of Computer Applications* 123: 32–38. [CrossRef]
Wang, Li, Haofei Zou, Jia Su, Ling Li, and Sohail Chaudhry. 2013. An arima-ann hybrid model for time series forecasting. *Systems Research and Behavioral Science* 30: 244–59. [CrossRef]
Wei, William W.S. 2006. *Time Series Analysis: Univariate and Multivariate Methods*. Time Series Analysis: Univariate and Multivariate Methods. Boston: Pearson Addison Wesley.
Zadeh, Lotfi A. 1965. Fuzzy sets. *Information and Control* 8: 338–53. [CrossRef]

risks

MDPI

Article

Foreign Exchange Futures Trading and Spot Market Volatility in Thailand

Woradee Jongadsayakul

Department of Economics, Faculty of Economics, Kasetsart University, Bangkok 10900, Thailand; fecowdj@ku.ac.th

Abstract: This paper investigates how the introduction of foreign exchange futures has an impact on spot volatility and considers the contemporaneous and dynamic relationship between spot volatility and foreign exchange futures trading activity, including trading volume and open interest in the Thailand Futures Exchange context, with the examples of the EUR/USD futures and USD/JPY futures. The results of the EGARCH (1,1) model show that the introduction of foreign exchange futures decreases spot volatility. It also increases the rate at which new information is impounded into spot prices but decreases the persistency of volatility shocks. A positive effect of unexpected trading volume and a negative effect of unexpected open interest on contemporaneous spot volatility are in line with the VAR(1) model results of the dynamic relationship between spot volatility and foreign exchange futures trading activity. With the impact on spot volatility caused by unexpected open interest rate being stronger than by unexpected trading volume, foreign exchange futures trading stabilizes spot volatility.

Keywords: foreign exchange futures; spot volatility; GARCH family models; VAR

1. Introduction

Since the 1997 Asian financial crisis, Thailand, South Korea, Indonesia, and the Philippines have all adopted a floating exchange rate regime, thereby increasing the importance of foreign exchange exposure management in East and Southeast Asia. The financial crisis originated in Thailand and caused the Thai government to float Thai baht on 2 July 1997. Facing volatile foreign exchange movements, exporters/importers, multinational companies, and overseas investment funds, have used financial derivatives for hedging exchange rate exposure. The value of foreign exchange (FX) derivatives activity has grown substantially over the last two decades. As illustrated in Figure 1, the FX derivatives volume traded in exchanges worldwide jumped from 46,947,055 contracts in 2000 to 990,925,534 contracts in 2009. After the 2007–2009 global financial crisis, trading of FX derivatives witnessed the biggest growth in volume in 2010, a 154.93 percent surge in yearly volume. This increase was driven mostly by Asian derivatives markets, which rose 239.84 percent and accounted for 75.81 percent of FX derivatives contracts traded on exchanges worldwide. Since the COVID-19 pandemic impacted global markets and volatility, the number of exchange-traded FX derivatives contracts has grown continuously to surpass the 4 billion mark in 2020. It reached the highest level in 2022, amounting to 7.73 billion contracts with 70.71 percent of them being traded in Asia. While trading of global FX derivatives decreased in volumes by 9.25 percent in 2023, the FX derivatives traded at Thailand Futures Exchange (TFEX) increased 12.19 percent compared to 2022, due to currency fluctuations and the popularity of FX trading. TFEX also launched new FX futures with EUR/USD and USD/JPY underlying on 31 October 2022. Although the most common hedging tools for Thai importers and exporters is a forward contract, Thai importers and exporters wishing to hedge their trade exposure may have limited access to forward contracts due to credit constraint or high transaction costs. They can use FX futures to better manage currency fluctuations.

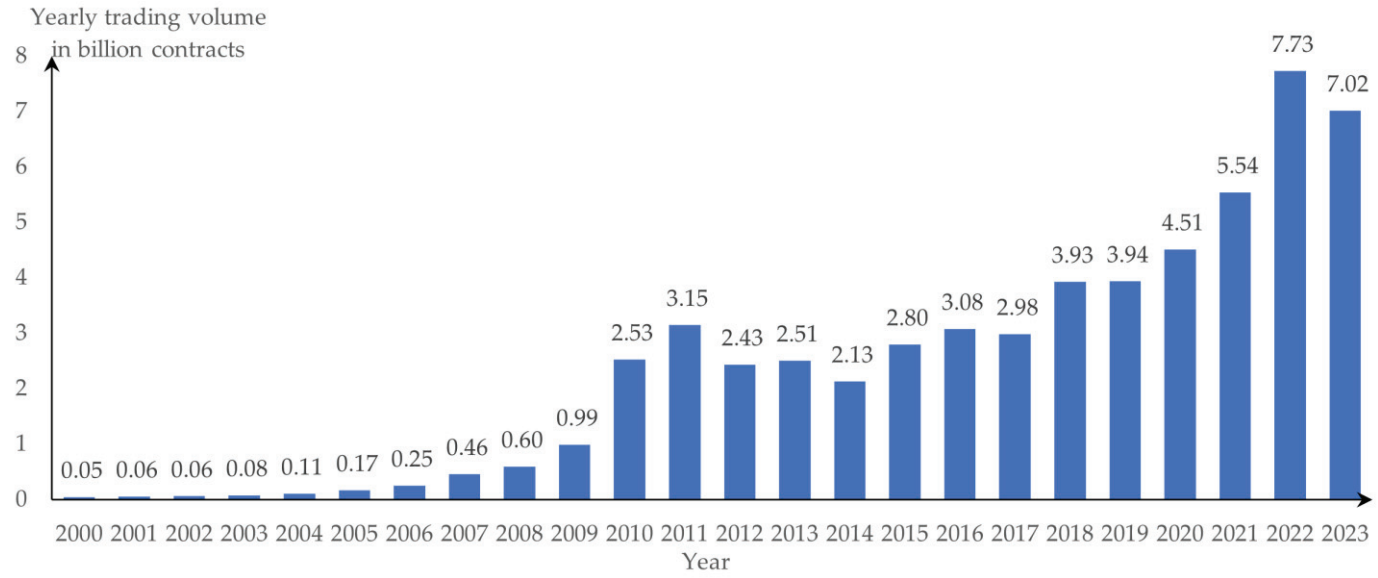


Figure 1. Volume of FX derivatives traded in exchanges worldwide from 2000 to 2023. Source: Futures Industry Association (2024).

Although the introduction of new FX futures provides traders with more options to match their investment goals and risk tolerance, it may cause an increase in the volatility of underlying exchange rates. The impact of FX futures trading on spot volatility has been widely investigated for major markets, but the empirical evidence is mixed. Some research, as detailed in the next section, shows that FX derivatives trading stabilizes the FX market by reducing its volatility. The FX derivatives market attracts additional traders to the underlying spot market, contributes to efficient price discovery, and leads to an increase in market depth. Other research, on the other hand, reveals that FX derivatives trading leads to an increase in spot volatility. This destabilizing impact on spot volatility is based on high leverage and speculative activities in the derivatives market. This study therefore aims to investigate the impact of FX derivatives trading on the volatility of the FX market in Thailand and to consider the contemporaneous and dynamic relationship between the volatility of the FX market and FX futures trading activity, with the examples of the EUR/USD futures and USD/JPY futures.

Using Generalized Autoregressive Conditional Heteroskedasticity (GARCH) family models augmented with dummy variable to investigate how the introduction of the EUR/USD futures and USD/JPY futures affects spot volatility, the empirical results show that the EGARCH (1,1) model is the best fitting model and highlight the evidence of the stabilizing effect of the introduction of FX futures on spot volatility. The launch of FX futures also increases the rate at which new information is incorporated into underlying spot prices and decreases the persistency of volatility shocks. In addition, the results show the destabilizing effect of unexpected trading volume and the stabilizing effect of unexpected open interest on contemporaneous spot volatility. These results are in line with the VAR(1) model results of the dynamic relationship between spot volatility and FX futures trading activity.

This study complements the literature about the stabilizing impact of FX derivatives trading on the volatility of the FX market in Thailand. It provides more insightful empirical evidence of how the introduction of FX futures changes the volatility structure of the underlying spot market and the relationship between spot volatility and the level of futures trading, including trading volume and open interest. Thus, the findings offer new insights for policy makers in relation to the economic usefulness of the derivatives market in emerging markets.

The remainder of this study is organized as follows. Section 2 provides a brief literature review on the impact of derivatives trading on spot volatility. Section 3 presents the data

and methodology used for the analysis. Section 4 discusses an empirical analysis of the impact of FX futures trading on the volatility of underlying exchange rates in Thailand, whilst the final section presents the conclusion.

2. Literature Review

Empirical studies on the impact of FX derivatives trading on spot volatility come with different methodologies. Much of the empirical research literature employs Generalized Autoregressive Conditional Heteroskedasticity (GARCH)-type models for modelling volatility of spot returns. For example, Gokcan (2000) and Szczygielski and Chipeta (2023) model the volatility of emerging stock market returns by employing GARCH family models. To analyze the impact of FX derivatives trading on spot volatility, the previous literature includes a dummy variable for the introduction of FX derivatives in the conditional variance equation. Some studies find the significance of the negative dummy coefficient, indicating the stabilizing effect of FX derivatives for these currency pairs CAD/USD, DEM/USD, JPY/USD, CHF/USD (Shastri et al. 1996), MXN/USD (Jochum and Kodres 1998), JPY/INR, and GBP/INR (Sakthivel et al. 2017). In addition, Oduncu (2011) shows a statistically significant negative relationship between the introduction of futures trading and the volatility of the Turkish currency market. The results also suggest that recent news plays a greater role, while old news plays a smaller role in determining the underlying spot market volatility as a consequence of the introduction of futures trading. Another group of research; however, finds that the introduction of FX derivatives destabilizes the FX market for EUR/INR (Gupta 2017; Sakthivel et al. 2017). The results by Shastri et al. (1996) and Sahu (2012) indicate a positive but insignificant impact of FX derivatives trading on volatility in spot exchange rates for GBP/USD and EUR/INR, respectively.

Some researchers divide full samples into two sub-periods, pre and post FX derivatives introduction periods, and use the GARCH volatility model to compare the underlying spot market volatility before and after the introduction of FX derivatives. For example, Jin et al. (2021) capture the effects of FX derivatives on the volatility of USD/INR, USD/RUB, and USD/ZAR by choosing a cutoff date to represent the beginning of a stable rise in FX futures trading and using the GARCH model for both pre and post cutoff periods. They have not found any strong evidence supporting the destabilizing effect of the FX derivatives market. However, in contrast with the study by Jin et al. (2021), Rani et al. (2022) and Singh and Patra (2022) show that the unconditional variance of the USD/INR exchange rate decreases in the post period. Rani et al. (2022) and Singh and Patra (2022) also investigated GBP, EUR, and JPY futures traded on the National Stock Exchange. They came to the same conclusion that the unconditional variance in the GBP/INR exchange rate decreases in the post period. For EUR and JPY, Singh and Patra (2022) found decreasing volatility in exchange rate returns in the post period, but Rani et al. (2022) witnessed opposite results.

The other way to investigate the impact of FX derivatives trading on spot volatility is by finding the relationship between spot volatility and FX derivatives trading variables, such as trading volume or open interest. Some researchers, such as Chatrath et al. (1996), found a positive relationship between spot volatility and FX futures trading volume. Bhargava and Malhotra (2007) used trading volume and open interest to separate hedgers from speculators and day traders. They suggest trading volume as a measure of speculating activities and open interest as a measure of hedging positions. Their main finding is that speculators and day traders destabilize the market for currency futures. Guru (2010), in contrast, found no causality either between trading volume and exchange rate volatility, or between open interest and exchange rate volatility in the case of USD/INR futures trading. In addition, Jochum and Kodres (1998) employed the Markov Switching Autogressive Conditional Heteroscedasticity (SWARCH) model, augmented with futures trading volume as an additional explanatory variable in the conditional variance equation. Their findings show no statistically significant influence from futures trading volume on the underlying spot market volatility in the cases of HUF and BRL.

Another group of researchers used a ratio of futures trading volume to open interest as a proxy for futures trading activity. Röthig (2004) investigated the relationship between spot volatility and FX futures trading activity by employing a vector autoregressive (VAR) system. The GARCH (1,1) model was chosen for the estimation of spot volatility of the five currencies, including AUD, CHF, CAD, JPY, and KRW. The results show that futures trading activity granger causes spot volatility for all currencies except KRW. In addition, spot volatility reacts to a shock in futures trading activity for all currencies except KRW. Moreover, Sharma (2011) conducts a granger causality test to examine the relationship between spot volatility and USD/INR futures trading activity. The results indicate a bidirectional causal relationship between spot volatility and FX futures trading activity. Comparing spot volatility before and after the introduction of FX futures, spot volatility increases after the FX futures introduction. Sivarajadhanavel et al. (2016) also calculated futures trading activity by dividing trading volume by open interest and including it in the conditional variance equation. The results of the augmented GARCH (1,1) model show that FX futures trading activity increases USD/INR exchange rate volatility.

Numerous empirical studies have applied the approach of Bessembinder and Sequin (1992) by decomposing either trading volume or open interest into expected and unexpected components. The expected and unexpected components are commonly obtained by using an Autoregressive Integrated Moving Average (ARIMA) model and included in the conditional variance equation of GARCH family models. Their findings suggest that trading different types of derivatives influences spot volatility differently. Most of them reveal a significant positive coefficient with respect to unexpected trading volume (e.g., (Bessembinder and Sequin 1992) for the S&P 500 index; (Fleming and Ostdiek 1999) for crude oil; (Kumar 2009) for soybean, maize, castor seed, guar seed, gold, silver, aluminium, copper, zine, crude oil, and natural gas; (Malhotra and Sharma 2016) for soya bean oil and crude palm oil; (Yilgor and Mebounou 2016) for the BIST-30 index; (Zhang et al. 2021) for bitcoin). This implies that an increase in unexpected trading volume as more information shocks leads to an increase in spot volatility. Expected trading volume is negatively related to spot volatility in the cases of S&P 500 index (Bessembinder and Sequin 1992), European real estate securities (Lee et al. 2014), mustard seed (Malhotra and Sharma 2016), and bitcoin (Zhang et al. 2021; Conlon et al. 2024); however, it is positively related to the volatility in other markets (e.g., (Kumar 2009) for silver, aluminium, copper, zine, and crude oil; (Malhotra and Sharma 2016) for mentha oil and soya oil). Regarding expected open interest, it has a negative impact on spot volatility in the cases of crude oil (Fleming and Ostdiek 1999), European real estate securities (Lee et al. 2014), soya oil, and mustard seed (Malhotra and Sharma 2016). These findings suggest that the derivatives market improves market depth and reduces the volatility of the underlying spot market. Several studies such as Bessembinder and Sequin (1992) and Shenbagaraman (2003) show the insignificant coefficient of unexpected open interest in explaining spot volatility, contradicting the findings of Malhotra and Sharma (2016) that higher unexpected open interest in mentha oil futures leads to an increase in spot volatility, and those of Fleming and Ostdiek (1999) that higher unexpected open interest in crude oil futures leads to a decrease in spot volatility. Another study by Kumar (2009) uses a VAR model to explain the dynamic relationship between spot volatility and the unexpected component of futures trading activity in the context of the Indian commodity derivatives market. The results show the significant causality running from unexpected trading volume to the spot volatility of all commodities. Except for natural gas, unexpected trading volume causes an increase in the underlying spot market volatility. The effect of unexpected open interest is positive for soybean, crude oil, and copper, but negative for maize, gold, silver, and aluminium.

The impact of the introduction of FX derivatives trading has been different in different markets and in most of the cases, the analysis has been conducted in the context of leading organized exchanges. The literature in the context of organized exchanges in emerging markets like Thailand is scarce. Due to the developing country's vulnerability to speculative attacks and adverse financial market development, the investigation about the potential

role of FX futures trading in developing countries' exchange rate stability is significant. Therefore, this study empirically investigates the impact of FX futures trading on spot volatility in Thailand.

3. Data and Methodology

3.1. Data

To examine the impact of the introduction of FX futures on spot volatility, daily data used in this paper consists of the exchange rates of EUR/USD and USD/JPY for the period from 27 September 2021 to 12 January 2024, covering the period before and after the introduction of EUR/USD and USD/JPY futures on 31 October 2022. For trading FX futures, Thailand Futures Exchange (TFEX) announced the launch of a night trading session during 6:50 p.m.–11:55 p.m. on 27 September 2021 and extended night session trading hours until 3:00 a.m. on 15 January 2024. Since the previous literature such as Jongadsayakul (2024) shows the impact of night trading sessions on volatility, this specific time period is chosen to avoid any possible impact on volatility. In addition, to analyze the relationship between spot volatility and FX futures trading activity, the daily trading volume and open interest data of EUR/USD futures and USD/JPY futures covering the period from 31 October 2022 to 12 January 2024 are used.

Consistent with previous studies, this study calculates the daily exchange rate returns for the currency pairs EUR/USD and USD/JPY by finding the first difference in the natural logarithms of the daily exchange rates, as shown in Equation (1).

$$R_t = \ln S_t - \ln S_{t-1}, \tag{1}$$

where S represents the daily exchange rate and R is the daily exchange rate return.

The daily exchange rate return series consists of 558 observations (over the whole observation period), of which 263 observations belong to the pre introduction period (27 September 2021–28 October 2022), and the remaining 295 observations belong to the post introduction period (31 October 2022–12 January 2024). In the case of the post introduction period, daily trading volume and open interest data of EUR/USD futures and USD/JPY futures were obtained from SETSMART.

This study begins with stationarity testing of all the return series (pre and post introduction periods and whole sample) as well as trading volume and open interest series via an Augmented Dickey–Fuller (ADF) test. The ADF test is performed to test the null hypothesis of a unit root (H_0: b = 0) against the alternative hypothesis of stationarity (H_a: b < 0). The Schwarz Information Criterion (SC) is used for the optimal lag selection. There are three cases, including a pure random walk, a random walk with intercept, and a random walk with intercept and linear time trend. The test equations of three cases are presented in Equations (2)–(4), respectively.

$$\Delta y_t = b y_{t-1} + \sum_{j=2}^{l} \gamma_j \Delta y_{t-j+1} + u_t, \tag{2}$$

$$\Delta y_t = c + b y_{t-1} + \sum_{j=2}^{l} \gamma_j \Delta y_{t-j+1} + u_t, \tag{3}$$

$$\Delta y_t = c + b y_{t-1} + \sum_{j=2}^{l} \gamma_j \Delta y_{t-j+1} + dt + u_t, \tag{4}$$

The descriptive statistics of daily exchange rate returns, trading volume, and open interest are presented in Table 1. It also contains the results of the unit root test and selected ARIMA(p,d,q) models.

Table 1. Descriptive statistics and ARIMA models of daily returns, trading volume, and open interest.

Series	n	Mean	Maximum	Minimum	Standard Deviation	ADF in Level [in 1st Difference] [1]	ARIMA(p,d,q) Model [2]
Panel A: EUR/USD							
Spot returns (whole sample)	558	−0.0114%	1.7211%	−1.8129%	0.5219%	−22.1842 ***	(0,0,0)
Spot returns (pre introduction)	263	−0.0605%	1.4397%	−1.8129%	0.5523%	−14.5942 ***	(0,0,0)
Spot returns (post introduction)	295	0.0323%	1.7211%	−1.4024%	0.4901%	−17.1121 ***	(0,0,0)
Futures volume	295	889.60	2897	54	580.11	−13.1272 ***	(1,0,1)
Open interest	295	1974.97	6197	75	1326.87	−2.4335 [21.5551 ***]	(0,1,1)
Panel B: USD/JPY							
Spot returns (whole sample)	558	0.0500%	2.5703%	−3.3618%	0.6406%	−18.7302 ***	(2,0,0)
Spot returns (pre introduction)	263	0.1095%	2.1451%	−3.3618%	0.5835%	−14.9389 ***	(0,0,0)
Spot returns (post introduction)	295	−0.0030%	2.5703%	−3.0036%	0.6842%	−14.3823 ***	(2,0,0)
Futures volume	295	3272.36	13,476	177	2133.69	−9.7128 ***	(2,0,1)
Open interest	295	7283.98	17,244	230	4676.60	−3.7239 **	(1,0,0)

Notes: Table 1 presents descriptive data statistics for EUR/USD (Panel A) and USD/JPY (Panel B). The whole samples for the EUR/USD and USD/JPY exchange rate returns range from 27 September 2021 to 12 January 2024, involving 558 observations. The whole samples are further classified into two sub-periods, pre introduction period (27 September 2021–28 October 2022) and post introduction period (31 October 2022–12 January 2024). The ADF test for stationarity is performed for spot returns, futures volume, and open interest. ** indicates significance at the 0.05 level, and *** indicates significance at the 0.01 level. All variables, except open interest in EUR/USD futures market, are stationary in levels. [1] With the presence of unit root test in level, test for unit root in 1st difference is conducted for open interest in EUR/USD futures market. [2] The correct ARIMA(p,d,q) model for the series depends on SC.

The results show that over the entire period from 27 September 2021 to 12 January 2024, the average daily exchange rate returns for the currency pairs EUR/USD and USD/JPY are −0.0114% and 0.05%, respectively. Throughout the entire period, the maximum (minimum) returns for the EUR/USD and USD/JPY exchange rates are 1.7211% (−1.8129%) and 2.5703% (−3.3618%), respectively. The standard deviations are 0.5219% for the EUR/USD exchange rate return and 0.6406% for the USD/JPY exchange rate return. A higher standard deviation indicates higher volatility in the FX market for the USD/JPY exchange rate compared to the EUR/USD exchange rate. The whole time period is divided into 2 sub-periods, pre introduction period (27 September 2021–28 October 2022) and post introduction period (31 October 2022–12 January 2024). The most volatile exchange rate is still USD/JPY during the pre and post introduction periods. The average daily returns of the EUR/USD exchange rate are negative during the pre introduction period and positive during the post introduction period, while those of the USD/JPY exchange rate witness the opposite results. Although both EUR/USD futures and USD/JPY futures were introduced on the same date, the USD/JPY futures contract is more actively traded than the EUR/USD futures contract in terms of trading volume and open interest. The mean USD/JPY futures trading volume is 3272 contracts, with an average open interest of 7284 contracts, while the mean EUR/USD futures trading volume is only 890 contracts, with an average open interest of 1975 contracts. In addition, at the 1% significance level, the ADF test for unit root in level rejects the presence of unit root in all the data series, except open interest. The EUR/USD futures open interest is non-stationary in level, but it is stationary after the 1st difference at

the 1% significance level. The USD/JPY futures open interest is stationary in level with a significance level of 5%. The ARIMA(p,d,q) models of all data series are discussed in Section 3.2.1.

3.2. Methodology

This study uses the following types of models for the analysis: (1) Autoregressive Integrated Moving Average (ARIMA) models, (2) Generalized Autoregressive Conditional Heteroskedasticity (GARCH) family models, and (3) Vector Autoregression (VAR) model.

3.2.1. ARIMA Models

In this section, ARIMA(p,d,q) models originally developed by Box and Jenkins (1976) are explored, where p denotes the number of autoregressive terms, d the number of times the series has to be differenced before it becomes stationary, and q the number of moving average terms. Therefore, an ARIMA(p,1,q) time series has to be differenced once (d = 1) before it becomes stationary, and the (first-differenced) stationary time series can be modelled as an ARMA(p,q) process. If d = 0, an ARIMA(p,0,q) process means ARMA(p,q), an ARIMA(p,0,0) process means a purely AR(p) stationary process, and an ARIMA(0,0,q) means a purely MA(q) stationary process. The equations for the ARMA(p,q) model, the AR(p) model, and the MA(q) model are stated in Equations (5)–(7), respectively.

$$\mathrm{ARMA(p,q)}: \; y_t = \mathrm{C} + \sum_{i=1}^{p} \lambda_i y_{t-i} + \sum_{i=0}^{q} \mu_i \varepsilon_{t-i};\; \mu_o = 1, |\lambda_i| < 1 \; (i = 1, \ldots, p) \text{ and } |\mu_i| < 1 \; (i = 1, \ldots, q), \tag{5}$$

$$\mathrm{AR(p)}: \; y_t = \mathrm{C} + \sum_{i=1}^{p} \lambda_i y_{t-i} + \varepsilon_t;\; |\lambda_i| < 1 \; (i = 1, \ldots, p), \tag{6}$$

$$\mathrm{MA(q)}: \; y_t = \mathrm{C} + \varepsilon_t + \sum_{i=1}^{q} \mu_i \varepsilon_{t-i};\; |\mu_i| < 1 \; (i = 1, \ldots, q), \tag{7}$$

The GARCH family models, as detailed in Section 3.2.2, are proposed to model the volatility of the underlying exchange rate. It is interesting to combine linear time series ARIMA with GARCH conditional variance. The ARIMA/GARCH model employs the ARIMA(p,d,q) model for the conditional mean and the GARCH family models for conditional variance. ARIMA captures the changes in the mean return, while GARCH presents the variance change in the residuals issued from the mean equation. As shown in Table 1, all exchange rate returns are stationary in levels. The condition mean equations are just constant mean equations for all EUR/USD returns series and for a series of USD/JPY returns during the pre introduction period. For the post introduction period and whole sample, the ARIMA(2,0,0) model was chosen for the USD/JPY returns series.

In addition, this research uses the ARIMA(p,d,q) models to decompose FX futures trading activity, including trading volume and open interest, into expected and unexpected components. The expected component is given by the model forecast, and the unexpected component is the difference between the actual and the fitted values. Table 1 shows the appropriate ARIMA models for trading volume and open interest in the FX futures market. For the EUR/USD futures, this study employs the ARIMA(1,0,1) model for trading volume and the ARIMA(0,1,1) model for open interest, while for the USD/JPY futures, this study chooses the ARIMA(2,0,1) model for trading volume and the ARIMA(1,0,0) model for open interest. The expected and unexpected components of FX futures trading activity are applied to the conditional variance of the selected GARCH model. It is interesting to include expected and unexpected components of FX futures trading activity as exogenous variables in volatility models and to examine the corresponding information. Only unexpected trading volume and unexpected open interest are included in the VAR model, as detailed in Section 3.2.3, to analyze the dynamic relationship between spot volatility and FX futures trading activity.

3.2.2. GARCH Family Models

This study adopts GARCH family models to examine spot volatility. Figures 2 and 3 are spot returns for the EUR/USD and USD/JPY exchange rates, respectively, which exhibit a volatility clustering property. The GARCH model proposed by Bollerslev (1986) was designed to capture the volatility clustering property in financial data (Jongadsayakul 2020, 2023). The GARCH (1,1) model is usually employed for modelling the volatility of a wide range of assets, as empirically demonstrated in previous studies (e.g., (Miaha and Rahmanb 2016) for stock; (Gokcan 2000) for stock; (Zhang et al. 2021) for bitcoin; (Jongadsayakul 2024) and for USD futures). However, the GARCH (1,1) model is a symmetric model with a non-negativity constraint on the parameters of the conditional variance. Since negative shocks are assumed to have a bigger impact on volatility than positive shocks in many financial markets, this study employs the asymmetric GARCH models, including the TARCH (1,1) model proposed by Zakoian (1990) and Glosten et al. (1993) and the EGARCH (1,1) model proposed by Nelson (1991). The EGARCH (1,1) model also ensures positive conditional variance without any restrictions on the parameters.

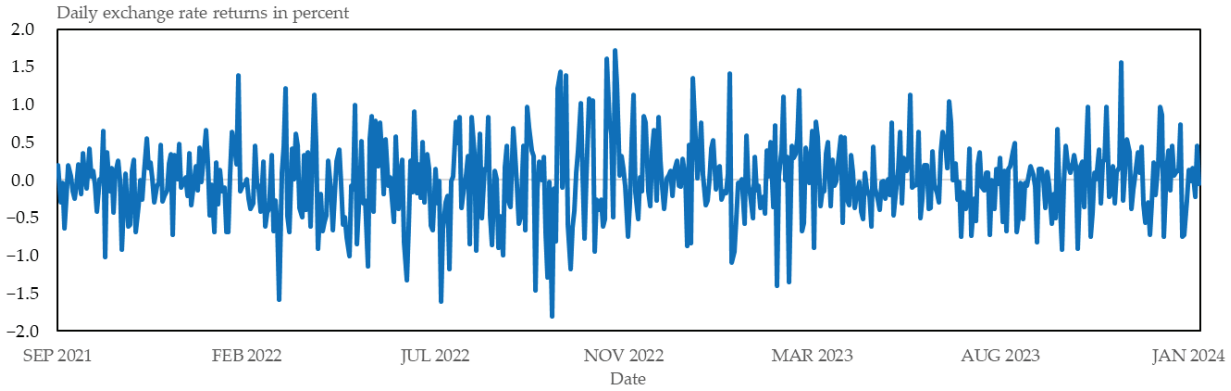


Figure 2. Daily spot returns of EUR/USD from 27 September 2021 to 12 January 2024.

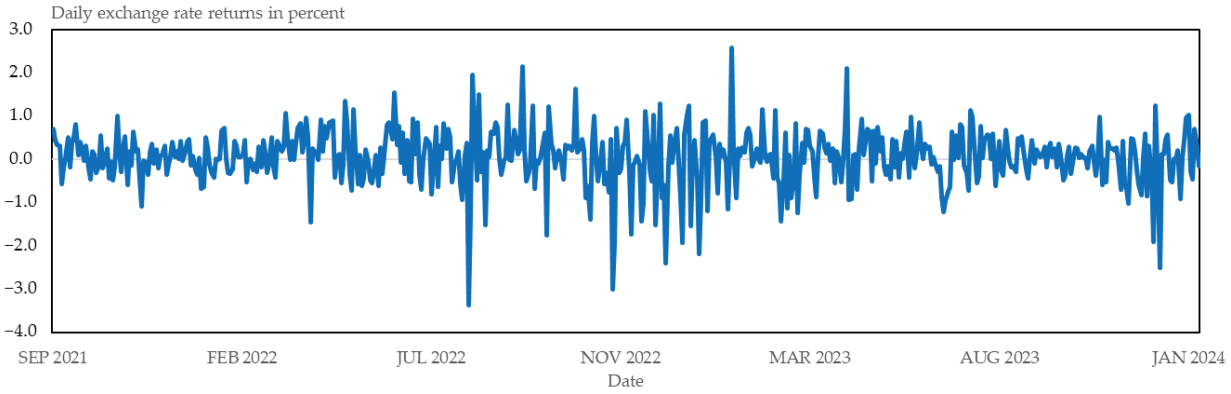


Figure 3. Daily spot returns of USD/JPY from 27 September 2021 to 12 January 2024.

To investigate the impact of the introduction of FX futures on spot volatility, this study adds a dummy variable (NEW), taking value 0 for the pre introduction period and 1 for the post introduction period, in the conditional variance equation. Thus, the conditional variance equations of the GARCH (1,1), TARCH (1,1), and EGARCH (1,1) models are shown in Equations (8)–(10), respectively.

$$\text{GARCH}\ (1,1):\ h_t^2 = \alpha_0 + \alpha_1 \varepsilon_{t-1}^2 + \beta_1 h_{t-1}^2 + aNEW_t; \alpha_i > 0, i = 0,1 \text{ and } \beta_1 > 0, \quad (8)$$

$$\text{TARCH}\ (1,1):\ h_t^2 = \alpha_0 + \alpha_1 \varepsilon_{t-1}^2 + \gamma \varepsilon_{t-1}^2 d_{t-1} + \beta_1 h_{t-1}^2 + aNEW_t, \quad (9)$$

$$\text{EGARCH}\ (1,1):\ \ln\left(h_t^2\right) = \alpha_0 + \alpha_1 \left|\frac{\varepsilon_{t-1}}{h_{t-1}}\right| + \gamma \frac{\varepsilon_{t-1}}{h_{t-1}} + \beta_1 \ln\left(h_{t-1}^2\right) + aNEW, \quad (10)$$

where h^2 is the conditional variance, ε^2 is square of residual, d_{t-1} takes a value of 0 if $\varepsilon_{t-1} \geq 0$ and a value of 1 if $\varepsilon_{t-1} < 0$, and $\frac{\varepsilon}{h}$ is standardized residual. For the existence of leverage effect, the sign of γ must be positive for Equation (9), such that the ARCH effect of $\alpha_1 + \gamma$ for bad news is larger than one of α_1 for good news. The negative sign of γ in Equation (10) indicates that the reaction to negative shocks $(\alpha_1 - \gamma)$ is larger than the reaction to positive shocks $(\alpha_1 + \gamma)$. To select the best volatility model of exchange rate returns for the whole sample, it is based on the lowest values of Akaike Information Criterion (AIC), Schwarz Criterion (SC), and Hannan Quinn (HQ).

In addition, the selected model without the dummy variable (NEW) in the conditional variance equation was estimated for the pre and post introduction periods. Comparing the ARCH and GARCH coefficients of the volatility model for the pre and post introduction periods provides more details about how FX futures trading has an impact on spot volatility and to what extent.

For the contemporaneous relationship between spot volatility and FX futures trading activity (trading volume and open interest), the expected and unexpected components of trading volume and those of open interest are included as explanatory variables in the conditional variance equation of the selected GARCH model for the post introduction period.

3.2.3. VAR Model

Using the VAR model as suggested by Kumar (2009), this research examines the dynamic interactions among spot volatility, the unexpected component of trading volume, and the unexpected component of open interest in the FX futures market for daily data between 31 October 2022 and 12 January 2024 (post introduction period). As discussed by Malhotra and Sharma (2016), information contained in the expected component of futures trading activity is already reflected in the spot prices. Therefore, only unexpected trading volume and unexpected open interest are used for the dynamic relationship analysis. The optimal lag order in the VAR model is then picked by considering the lowest values of AIC, SC, and HQ. The VAR(p) model can be written as follows:

$$h_t^2 = a_1 + \sum_{j=1}^{p} b_{1j} h_{t-j}^2 + \sum_{j=1}^{p} b_{2j} UV_{t-i} + \sum_{j=1}^{p} b_{3j} UO_{t-i} + e_{1t}, \tag{11}$$

$$UV_t = a_2 + \sum_{j=1}^{p} c_{1j} h_{t-j}^2 + \sum_{j=1}^{p} c_{2j} UV_{t-i} + \sum_{j=1}^{p} c_{3j} UO_{t-i} + e_{2t}, \tag{12}$$

$$UO_t = a_3 + \sum_{j=1}^{p} d_{1j} h_{t-j}^2 + \sum_{j=1}^{p} d_{2j} UV_{t-i} + \sum_{j=1}^{p} d_{3j} UO_{t-i} + e_{3t}, \tag{13}$$

where h^2 is the estimated conditional variance obtained from the selected GARCH model, *UV* is the unexpected component of FX futures trading volume obtained from the selected ARIMA model, and *UO* is the unexpected component of open interest obtained from the selected ARIMA model.

The Granger causality test, variance decomposition, and impulse response function are conducted while analyzing this dynamic relationship.

4. Results and Discussion

This section presents the results of the empirical analysis and provides some discussion.

4.1. Effect of FX Futures Introduction on Spot Volatility

Table 2 shows the estimation results of the GARCH family models, the GARCH (1,1), TARCH (1,1), and EGARCH (1,1) models, augmented with the dummy variable (NEW) for the period from 27 September 2021 to 12 January 2024 (whole sample) for the currency pairs EUR/USD and USD/JPY. As mentioned earlier, this study adds the dummy variable (NEW) in the conditional variance equation to analyze the effect of the

introduction of new FX futures on spot volatility. All estimated models are first checked for appropriateness. The *p*-values from the Ljung–Box Q tests on the standardized residuals and squared residuals as well as the Lagrange Multiplier (LM) test for the remaining ARCH effects in the standardized residuals are greater than the 5% significance level. The insignificant Q statistics provide evidence for failing to reject the null hypothesis of no autocorrelation in the estimated GARCH family models' residuals. The insignificant LM test statistics show no sign of additional ARCH effects left in the standardized residuals. Thus, the use of GARCH (1,1), TARCH (1,1), and EGARCH (1,1) models for modelling the volatility of exchange rates for EUR/USD and USD/JPY is appropriate. In addition, as indicated in Table 1, the analysis applies the constant mean equation for EUR/USD returns and the AR(2) specification for the mean equation of USD/JPY returns. The AR terms are all significant at the 1% significance level for the GARCH (1,1) and EGARCH (1,1) models, and at the 5% significance level for the TARCH (1,1) model.

Table 2. Estimation results of the GARCH family models for whole sample.

Exchange Rate		**EUR/USD**			**USD/JPY**	
Model	**GARCH (1,1)**	**TARCH (1,1)**	**EGARCH (1,1)**	**GARCH (1,1)**	**TARCH (1,1)**	**EGARCH (1,1)**
Panel A: Mean equation						
Constant (C)	-0.0002	-0.0002	-0.0002	0.0006	0.0006	0.0004
	(0.4104)	(0.4050)	(0.3258)	(0.0139) **	(0.0255) **	(0.0967) *
AR(1) (λ_1)				0.1181	0.1114	0.1204
				(0.0064) ***	(0.0114) **	(0.0048) ***
AR(2) (λ_2)				-0.1290	-0.1183	-0.1188
				(0.0064) ***	(0.0121) **	(0.0091) ***
Panel B: Variance equation						
Constant (α_0)	2.36×10^{-7}	2.33×10^{-7}	-0.0527	3.55×10^{-7}	5.92×10^{-7}	-0.1239
	(0.0099) ***	(0.0001) ***	(0.0010) ***	(0.0458) **	(0.0064) ***	(0.0047) ***
ARCH (α_1)	-0.0105	-0.0096	0.0024	0.0491	0.0276	0.0987
	(0.1203)	(0.0670) *	(0.8484)	(0.0000) ***	(0.1658)	(0.0000) ***
Asym. (γ)		0.0020	-0.0110		0.0385	-0.0358
		(0.8589)	(0.5454)		(0.0823) *	(0.0414) **
GARCH (β_1)	1.0066	1.0048	0.9947	0.9482	0.9446	0.9940
	(0.0000) ***	(0.0000) ***	(0.0000) ***	(0.0000) ***	(0.0000) ***	(0.0000) ***
NEW (a)	-2.09×10^{-7}	-2.08×10^{-7}	-0.0081	-1.67×10^{-7}	-3.51×10^{-7}	-0.0161
	(0.0000) ***	(0.0000) ***	(0.0617) *	(0.2162)	(0.0316) **	(0.0011) ***
Panel C: Residual diagnostics						
Q(36)	42.568	42.116	39.484	33.313	34.682	33.439
	(0.209)	(0.223)	(0.317)	(0.501)	(0.435)	(0.495)
$Q^2(36)$	27.261	27.020	30.586	26.738	27.219	32.592
	(0.852)	(0.860)	(0.724)	(0.869)	(0.854)	(0.631)
ARCH-LM (1)	2.1333	1.8827	0.6311	0.2027	0.1651	0.7424
	(0.1441)	(0.1700)	(0.4269)	(0.6526)	(0.6845)	(0.3889)
Panel D: Model selection						
AIC	-7.7504	-7.7454	-7.7415	-7.3962	-7.3961	-7.4042
SC	-7.7117	-7.6989	-7.6950	-7.3418	-7.3340	-7.3420
HQ	-7.7353	-7.7272	-7.7234	-7.3749	-7.3718	-7.3799

Notes: Table 2 shows the estimation results of the GARCH family models augmented with the dummy variable (NEW) for the introduction of FX futures. *p*-values are in parentheses with the use of *, **, and *** to indicate significance levels at 0.10, 0.05, and 0.01, respectively. The dummy variable coefficient (a) is negative, indicating the stabilizing effect of the introduction of FX futures on spot volatility.

For the case of EUR/USD, although the GARCH (1,1) and TARCH (1,1) models have lower values of AIC, SC, and HQ than the EGARCH (1,1) model, their ARCH coefficients (α_1) are negative (though insignificant in the GARCH (1,1) model), violating the condition of non-negativity on the parameters of the conditional variance equation. Therefore, the EGARCH (1,1) without any restrictions on the parameters was chosen for modelling the volatility of EUR/USD returns. The estimation result of the EGARCH (1,1) model reveals the insignificant coefficient of ARCH term (α_1), meaning that recent news does not have an impact on changes in the EUR/USD exchange rate. However, the significant coefficient of GARCH term (β_1) confirms the effect of old news on the volatility of EUR/USD returns. In addition, the leverage effect does not exist in the FX market for EUR/USD due to the insignificance of the asymmetric coefficient (γ). The results further show that the introduction of EUR/USD futures reduces spot volatility since the coefficient of the dummy variable (NEW) is negative and statistically significant at the 10% level.

For the case of USD/JPY, the GARCH (1,1) and TARCH (1,1) models satisfy the condition of non-negativity on the parameters of the conditional variance equation. Both TARCH (1,1) and EGARCH (1,1) models show the existence of leverage effect in the FX market for USD/JPY due to the positive sign of γ in the TARCH (1,1) model and the negative sign of γ in the EGARCH (1,1) model. However, the EGARCH (1,1) model is best suited based on its lowest values of AIC, SC, and HQ. The estimation result of the EGARCH (1,1) model reveals the significant coefficients of the ARCH term (α_1) and the GARCH term (β_1), meaning that recent news and past news have an impact on spot volatility. In addition, the introduction of USD/JPY futures reduces spot volatility since the coefficient of the dummy variable (NEW) is negative and statistically significant at the 1% level.

Therefore, the estimated coefficient on the dummy variable (NEW) is negative and significant in the EGARCH (1,1) model of exchange rate returns for EUR/USD and USD/JPY, implying that the introduction of new FX futures by TFEX results in a decrease in spot volatility. This result is consistent with the existing literature (see for example, Shastri et al. 1996; Jochum and Kodres 1998; Oduncu 2011; Sakthivel et al. 2017), which provides evidence of the stabilizing effect of the introduction of FX derivatives on spot volatility.

4.2. Pre and Post Introduction Comparison of Spot Volatility and Contemporaneous Relationship between Spot Volatility and FX Futures Trading

The whole sample was divided into two sub-periods, the pre introduction period (27 September 2021–28 October 2022), and the post introduction period (31 October 2022–12 January 2024). The EGARCH (1,1) model without the dummy variable (NEW) was chosen for modelling the volatility of EUR/USD and USD/JPY returns. To understand how and to what extent the introduction of FX futures affects spot volatility, the ARCH and GARCH coefficients for the pre introduction period were compared with those for the post introduction period. In addition, to investigate the contemporaneous relationship between spot volatility and FX futures trading activity (trading volume and open interest), the EGARCH (1,1) model was extended by adding FX futures trading activity in the condition variance equation for the post introduction period. Trading activity in the FX futures market, including trading volume and open interest, can be decomposed of expected and unexpected components using the ARIMA(p,d,q) model. Table 3 represents the outcomes of the EGARCH (1,1) model for pre and post introduction periods as well as the outcomes of the extended EGARCH (1,1) model with FX futures trading activity for post introduction period.

As shown in Table 3, the EGARCH (1,1) model and its extension are appropriate for modelling the volatility of the EUR/USD and USD/JPY returns based on diagnostic tests for serial correlation and remaining ARCH effect. All *p*-values from the Ljung–Box Q tests on the standardized residuals and squared residuals and the ARCH-LM test are greater than the 5% significance level, indicating no evidence of serial correlation and remaining ARCH effect. With the selected ARIMA(p,d,q) models shown in Table 1, the EGARCH (1,1) model of EUR/USD returns is estimated with the constant mean equation for both the pre

and post introduction periods. For USD/JPY returns, the analysis applies the EGARCH (1,1) model with the constant mean equation for the pre introduction period and that with the AR(2) for the post introduction period.

Table 3. Estimation results of the EGARCH (1,1) and extended EGARCH (1,1) models.

Exchange Rate		**EUR/USD**			**USD/JPY**	
Period	**Pre Intro.**	**Post Intro.**	**Post Intro.**	**Pre Intro.**	**Post Intro.**	**Post Intro.**
			Panel A: Mean equation			
Constant (C)	−0.0010	0.0003	0.0002	0.0013	0.0005	0.0003
	(0.0002) ***	(0.3706)	(0.4380)	(0.0001) ***	(0.1416)	(0.3369)
AR(1) (λ_1)					0.0450	0.0601
					(0.3920)	(0.2793)
AR(2) (λ_2)					−0.1622	−0.1496
					(0.0051) ***	(0.0096) ***
			Panel B: Variance equation			
Constant (α_0)	−0.0856	−9.9206	−10.1623	0.0085	−1.1219	−1.3627
	(0.0365) **	(0.0030) ***	(0.0000) ***	(0.9208)	(0.0004) ***	(0.0013) ***
ARCH (α_1)	−0.0330	0.2912	0.3467	−0.0109	0.1883	−0.0824
	(0.4135)	(0.0490) **	(0.0247) **	(0.6891)	(0.0210) **	(0.2746)
Asym. (γ)	−0.0375	0.1629	0.1813	0.0570	−0.2347	−0.1854
	(0.1423)	(0.0815) *	(0.0520) *	(0.0092) ***	(0.0000) ***	(0.0042) ***
GARCH (β_1)	0.9885	0.0921	0.0819	0.9993	0.9049	0.8468
	(0.0000) ***	(0.7664)	(0.6503)	(0.0000) ***	(0.0000) ***	(0.0000) ***
Expected vol.			−0.00001			−0.000004
			(0.9853)			(0.8587)
Unexpected vol.			0.0008			0.0002
			(0.0001) ***			(0.0000) ***
Expected OI			0.0006			−0.00002
			(0.5241)			(0.0167) **
Unexpected OI			−0.0007			−0.0001
			(0.0209) **			(0.0813) *
			Panel C: Residual diagnostics			
Q(36)	25.673	44.747	39.639	42.384	28.985	32.009
	(0.899)	(0.150)	(0.311)	(0.215)	(0.712)	(0.566)
$Q^2(36)$	31.286	27.917	34.992	35.418	26.783	24.268
	(0.692)	(0.830)	(0.516)	(0.496)	(0.868)	(0.932)
ARCH-LM (1)	0.0632	0.0790	0.0001	1.9487	0.1552	0.1847
	(0.8015)	(0.7786)	(0.9910)	(0.1627)	(0.6936)	(0.6674)

Notes: Table 3 shows the EGARCH (1,1) estimation results for pre and post introduction periods and the estimation results of the EGARCH (1,1) model augmented with futures trading activity for post introduction period. *p*-values are in parentheses with the use of *, **, and *** to indicate significance levels at 0.10, 0.05, and 0.01, respectively. The negative coefficient γ shows the existence of the leverage effect in the USD/JPY spot market after the introduction of USD/JPY futures. The coefficient of unexpected trading volume is positive and significant, indicating the destabilizing effect of unexpected trading volume on spot volatility. On the other hand, the coefficient of unexpected open interest is negative and significant, indicating the stabilizing effect of unexpected open interest on spot volatility.

For the case of EUR/USD, the results show an increase in the ARCH coefficient (α_1) for the post introduction period, suggesting a greater impact of recent news on changes in EUR/USD after the introduction of EUR/USD futures. On the other hand, there is a reduction in the GARCH coefficient (β_1), implying a decrease in the persistency of volatility shocks after the introduction of EUR/USD futures. In addition, the asymmetric coefficient (γ) becomes positive and is statistically significant at the 10% level after the introduction of EUR/USD futures, suggesting the presence of an asymmetric effect in the EUR/USD spot market after the introduction of EUR/USD futures. The EGARCH (1,1) model is also augmented with the expected and unexpected components of FX futures trading activity (trading volume and open interest). The extended EGARCH (1,1) model

confirms the presence of asymmetric effect in the EUR/USD spot market during the post introduction period due to the positive and significant coefficient γ at the 10% level. In addition, none of the coefficients on expected trading volume and expected open interest are statistically significant. Unlike the results from the expected component part, the coefficients of unexpected trading volume and unexpected open interest are significant. The positive coefficient of unexpected trading volume suggests an increase in spot volatility as a result of information shocks. Information shocks are expected to move prices and cause a sudden increase in volume in both underlying spot and futures markets. However, the negative coefficient of unexpected open interest suggests a decrease in the volatility of EUR/USD returns due to open interest shocks.

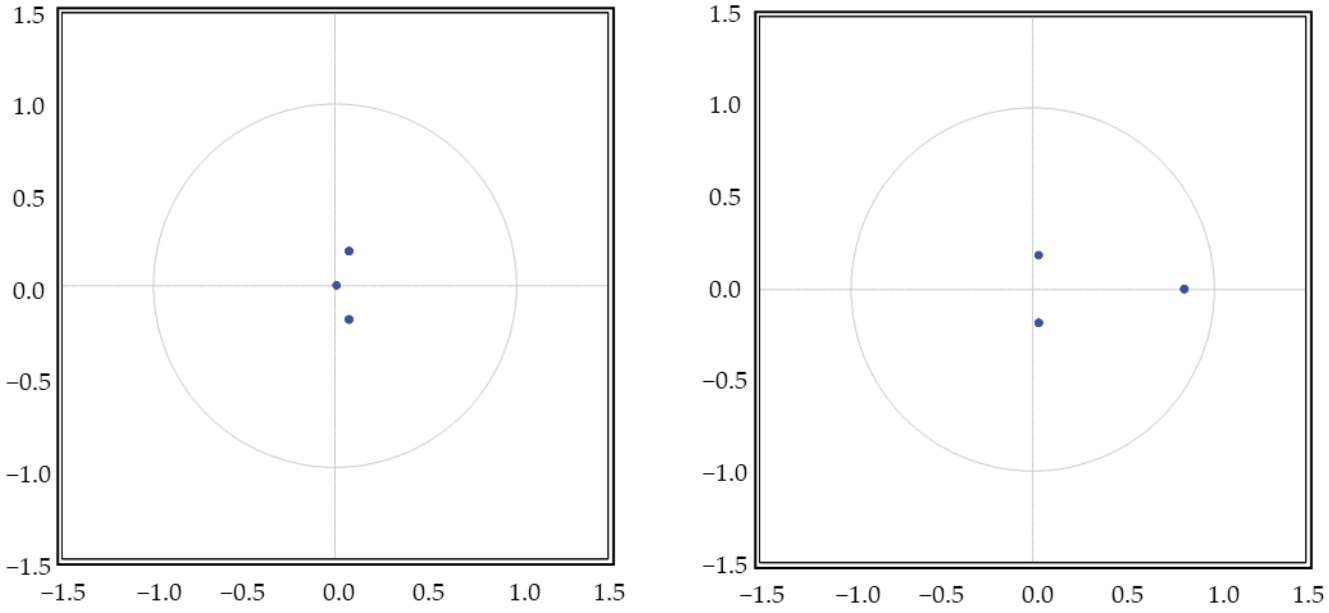
For the case of USD/JPY, recent news significantly influences changes in USD/JPY for the post introduction period, as the coefficient of ARCH term (α_1) is significant at the 5% level. An increase in the ARCH coefficient (α_1) for the post introduction period implies that recent news has a greater impact on changes in USD/JPY after the introduction of USD/JPY futures. In contrast, the GARCH coefficient (β_1) reflects the persistence of the effect of old news on volatility. The GARCH coefficient (β_1) is statistically significant at the 1% level for both the pre and post introduction periods. Its value falls after the introduction of USD/JPY futures, implying that old news has a lower impact on the current volatility of USD/JPY returns. In addition, the asymmetric coefficient (γ) is statistically significant at the 1% level for both the pre and post introduction periods. Its value becomes negative after the introduction of USD/JPY futures, suggesting the presence of a leverage effect in the USD/JPY spot market after the introduction of USD/JPY futures. Then, the expected and unexpected components of trading volume and those of open interest are included as additional explanatory variables in the EGARCH (1,1) model. The estimation result of the extended EGARCH (1,1) model confirms the existence of leverage effect in the USD/JPY spot market after the introduction of USD/JPY futures due to the negative and significant coefficient γ at the 1% level. In addition, only the coefficient of the expected trading volume is insignificant. The positive and significant coefficient of unexpected trading volume suggests that a sudden change in USD/JPY futures trading volume increases the volatility of USD/JPY returns. Furthermore, the negative and significant coefficients of the expected and unexpected components of open interest imply that lower volatility shocks are associated with a given volume in deeper markets, as suggested by Smit and Louw (1996). Smit and Louw (1996) discuss that the expected open interest component reflects open interest at the start of a trading day, while the unexpected open interest component reflects any change in open interest during a day. The negative coefficient of expected open interest suggests that the USD/JPY futures market improves market depth and stabilizes the USD/JPY spot market. The negative coefficient of unexpected open interest suggests the unanticipated daily change in open interest as a proxy for traders' willingness to risk capital mitigates volatility of the USD/JPY spot market.

The analyses for both EUR/USD and USD/JPY achieve the same results of a higher ARCH coefficient and a lower GARCH coefficient after the FX futures introduction. The introduction of FX futures increases the rate at which new information is incorporated into underlying spot prices and decreases the persistency of volatility shocks. This implies an improvement in efficiency in the underlying spot market as a result of the introduction of new FX futures by TFEX. This finding offers some empirical support for the presence of a stabilizing effect of futures trading on the underlying spot market. In addition, the results of the destabilizing effect of unexpected trading volume on volatility in spot exchange rates for EUR/USD and USD/JPY are in line with a large number of previous studies (e.g., Bessembinder and Sequin 1992; Fleming and Ostdiek 1999; Kumar 2009; Malhotra and Sharma 2016; Yilgor and Mebounou 2016; Zhang et al. 2021). The insignificant coefficient of expected trading volume is evident in both EUR/USD and USD/JPY, indicating the minor role of this variable on spot volatility. Consistent with the research on the crude oil market by Fleming and Ostdiek (1999), there exists a negative contemporaneous relationship between spot volatility and unexpected open interest in the EUR/USD and USD/JPY futures

markets. While the expected open interest as a proxy for market depth has a stabilizing influence on the USD/JPY spot market, its insignificant coefficient in the EUR/USD case may be attributed to lower trading activity, as discussed by Lee et al. (2014) in the European real estate case.

4.3. Dynamic Relationship between Spot Volatility and FX Futures Trading Activity

This section investigates the dynamic relationship between spot volatility and trading activity in the EUR/USD and USD/JPY futures markets by adopting a Vector Autoregressive (VAR) model. The VAR(1) model is chosen for both cases (EUR/USD and USD/JPY) based on the optimal lag order that minimizes the AIC, SC, and HQ values. In the VAR(1) model, the estimated conditional variance (h^2) obtained from the EGARCH (1,1) model and the unexpected components of trading volume (UV) and open interest (UO) obtained from the ARIMA models are stationary at the 1% level of significance. The stability test of the VAR(1) model was conducted as shown in Figure 4. Since all the inverse roots of the model have roots with a modulus less than one and lie inside the unit circle, the VAR(1) model is variance and covariance stationary.



(**a**) EUR/USD (**b**) USD/JPY

Figure 4. Inverse roots of AR characteristic polynomial. Notes: Figure 4 shows the inverse roots of the VAR(1) model for EUR/USD (**a**) and USD/JPY (**b**). The VAR(1) model for each case is stable since all roots have a modulus less than one and lie inside the unit circle.

Table 4 presents the estimated results of the VAR(1) model and the LM test results for autocorrelation. This study conducts an LM test of the null hypothesis of no autocorrelation for the first two lags of the residuals. All *p*-values associated with the LM test statistics are greater than 5%, meaning that there is no autocorrelation left in the residuals at lags 1 and 2. In addition, the estimation results of the VAR(1) model show that volatility of the EUR/USD and USD/JPY returns are positively affected by one day lag of unexpected trading volume (at the 1% significance level) and negatively affected by one day lag of unexpected open interest in the FX futures market for EUR/USD (at the 1% significance level) and USD/JPY (at the 5% significance level). The unexpected FX futures trading volume is found to be positively affected by one day lag of unexpected open interest in the FX futures market for EUR/USD (at the 10% significance level) and USD/JPY (at the 5% significance level). Regarding unexpected open interest in the FX futures market, it is positively affected by its lagged value at the significance level of 0.05 and negatively

affected by one day lag of unexpected trading volume at the significance level of 0.01. These results are confirmed through the Granger causality test (Table 5). The Granger causality test results show a significant causality running from unexpected trading volume to spot volatility or unexpected open interest to spot volatility. There exists a bi-directional relationship between unexpected trading volume and unexpected open interest in the FX futures market.

Table 4. Results of the estimated VAR(1) model and the LM test.

Exchange Rate	**EUR/USD**			**USD/JPY**		
Variables	h^2	UV	UO	h^2	UV	UO
Constant	0.00002	58.3469	10.8041	0.000008	73.6427	118.2093
	(0.0000) ***	(0.4959)	(0.8658)	(0.0001) ***	(0.6314)	(0.2485)
$h^2(-1)$	0.0606	−2,419,031	−460,694	0.8358	−844,566	−1,427,253
	(0.2919)	(0.4626)	(0.8513)	(0.0000) ***	(0.7100)	(0.3464)
$UV(-1)$	3.62×10^{-9}	−0.0200	−0.2213	2.38×10^{-9}	−0.0677	−0.2333
	(0.0015) ***	(0.7586)	(0.0000) ***	(0.0094) ***	(0.2946)	(0.0000) ***
$UO(-1)$	-4.29×10^{-9}	0.1471	0.1283	-2.66×10^{-9}	0.1971	0.1388
	(0.0040) ***	(0.0838) *	(0.0433) **	(0.0420) **	(0.0334) **	(0.0247) **
LM(1)		11.7995 (0.2249)			7.6433 (0.5705)	
LM(2)		13.2414 (0.1520)			10.8183 (0.2884)	

Notes: Table 4 shows the estimation results of the VAR model with the optimal lag order (1). *p*-values are in parentheses with the use of *, **, and *** to indicate significance levels at 0.10, 0.05, and 0.01, respectively. While one day lag of unexpected trading volume (UV) positively affects volatility of EUR/USD and USD/JPY returns, one day lag of unexpected open interest (UO) has a greater negative impact.

Table 5. Granger causality results.

Exchange Rate	**EUR/USD**		**USD/JPY**	
Null Hypothesis	**Chi-Square Statistic**	***p*-Value**	**Chi-Square Statistic**	***p*-Value**
UV does not Granger-cause h^2	10.1461	0.0014 ***	6.7722	0.0093 ***
UO does not Granger-cause h^2	8.3470	0.0039 ***	4.1494	0.0416 **
h^2 does not Granger-cause UV	0.5400	0.4624	0.1383	0.7099
UO does not Granger-cause UV	2.9974	0.0834 *	4.5409	0.0331 **
h^2 does not Granger-cause UO	0.0352	0.8512	0.8877	0.3461
UV does not Granger-cause UO	20.7472	0.0000 ***	29.2968	0.0000 ***

Note: Table 5 shows VAR Granger causality results. * indicates significance level at the 0.10 level, ** indicates significance level at the 0.05 level, and *** indicates significance level at the 0.01 level. There is evidence of bi-directional causality between unexpected trading volume (UV) and unexpected open interest (UO). Any changes in unexpected trading volume (UV) and unexpected open interest (UO) also affect volatility of EUR/USD and USD/JPY returns.

As discussed by Kumar (2009), the analysis on variance decomposition (Table 6) and impulse response function (Figure 5) exhibits information beyond the results of the VAR estimation and Granger causality tests. The percentage of variation in spot volatility explained by unexpected trading volume is about 2% for both the EUR/USD and USD/JPY cases. Unexpected open interest explains only 1% of the variation in the volatility of USD/JPY returns and about 2.55% of the variation in the volatility of EUR/USD returns. Contrarily, spot volatility explains less than 1% variation in the unexpected component of futures trading activity (trading volume and open interest) for both the EUR/USD and USD/JPY cases. These results are consistent with the Granger causality test results.

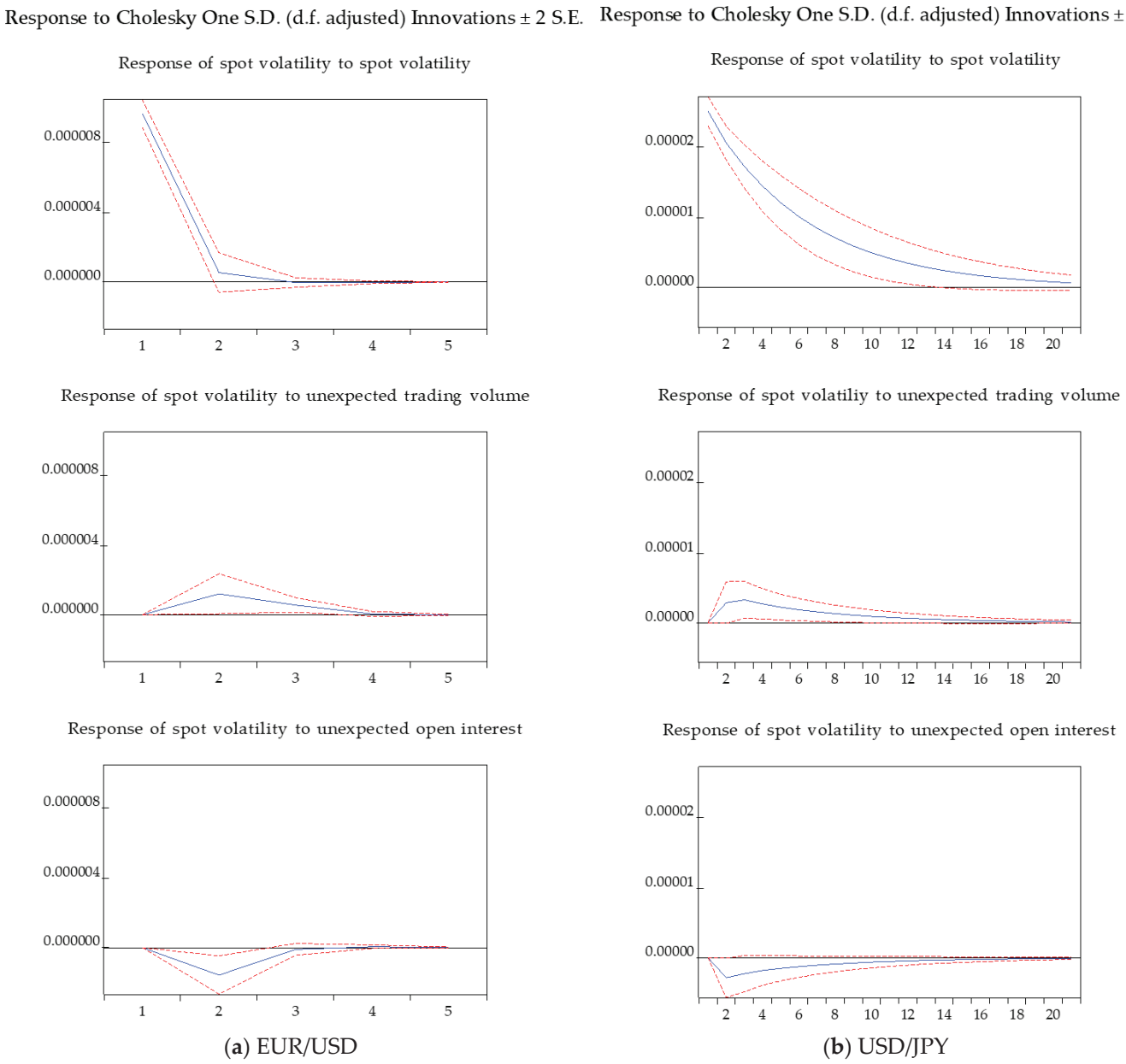
Table 6. Results of variance decomposition.

(%)	Variance Decomposition of h^2			Variance Decomposition of UV			Variance Decomposition of UO		
Period	h^2	UV	UO	h^2	UV	UO	h^2	UV	UO
				Panel A: EUR/USD					
1	100.0	0.00	0.00	0.00	100.0	0.00	0.08	20.80	79.12
2	95.99	1.47	2.54	0.15	98.91	0.94	0.08	24.74	75.18
3	95.68	1.78	2.54	0.15	98.88	0.97	0.09	24.86	75.06
4	95.68	1.78	2.54	0.15	98.88	0.97	0.09	24.86	75.06
5	95.68	1.78	2.55	0.15	98.88	0.97	0.09	24.86	75.06
6	95.68	1.78	2.55	0.15	98.88	0.97	0.09	24.86	75.06
7	95.68	1.78	2.55	0.15	98.88	0.97	0.09	24.86	75.06
8	95.68	1.78	2.55	0.15	98.88	0.97	0.09	24.86	75.06
9	95.68	1.78	2.55	0.15	98.88	0.97	0.09	24.86	75.06
10	95.68	1.78	2.55	0.15	98.88	0.97	0.09	24.86	75.06
				Panel B: USD/JPY					
1	100.0	0.00	0.00	0.13	99.87	0.00	0.51	18.86	80.63
2	98.49	0.76	0.75	0.13	98.49	1.37	0.47	24.77	74.77
3	97.70	1.35	0.95	0.14	98.48	1.38	0.53	24.85	74.62
4	97.37	1.61	1.02	0.15	98.46	1.38	0.56	24.84	74.59
5	97.20	1.75	1.06	0.16	98.46	1.38	0.59	24.84	74.58
6	97.09	1.83	1.08	0.17	98.45	1.38	0.61	24.83	74.56
7	97.02	1.88	1.09	0.17	98.45	1.38	0.62	24.83	74.55
8	96.98	1.92	1.10	0.17	98.44	1.38	0.63	24.83	74.55
9	96.95	1.94	1.11	0.18	98.44	1.38	0.63	24.83	74.54
10	96.93	1.95	1.11	0.18	98.44	1.38	0.64	24.82	74.54

Notes: Table 6 shows a forecast error variance decomposition analysis for both EUR/USD (Panel A) and USD/JPY (Panel B) cases. About 96 percent of the variation in spot volatility can be traced back to its own innovation, while the rest is explained by unexpected trading volume (UV) and unexpected open interest (UO).

In addition, this paper uses the impulse response function to analyze the response of spot volatility to a one standard deviation shock in unexpected trading volume and unexpected open interest for both the EUR/USD and USD/JPY cases. As shown in Figure 5, the response of spot volatility to its own shock is positive and high. It diminishes quickly in the case of EUR/USD and reaches equilibrium within 3 observation days, while in the case of USD/JPY, it exponentially decreases and reaches equilibrium within 21 observation days. For both the EUR/USD and USD/JPY cases, the response of spot volatility to shock in unexpected trading volume is positive, while spot volatility adjustment to shock in unexpected open interest is negative. It takes a few days (3–4 days) to die out in the case of EUR/USD, while it takes longer time (12–15 days) to die out in the case of USD/JPY. Therefore, these results are consistent with the Granger causality test results and are convincing proof of the sign results of the VAR(1) model.

Therefore, the analysis of the dynamic relationship between spot volatility and trading activity (unexpected component of trading activity) in the FX futures market for EUR/USD and USD/JPY provides evidence of the destabilizing impact of trading volume and the stabilizing impact of open interest on spot volatility. As discussed by Malhotra and Sharma (2016), futures market and spot market are linked by arbitrage. If there is an increase in unexpected futures trading volume driven by uninformed speculators, then spot volatility will increase. On the other hand, open interest is a measure of the positions of hedgers or actively informed traders who bring fundamental information to the futures market. The increase in hedging positions increases the market depth and consequently reduces spot volatility. Since unexpected open interest has a stronger influence on spot volatility than unexpected trading volume, FX futures trading stabilizes volatility in spot exchange rates for EUR/USD and USD/JPY.



(**a**) EUR/USD (**b**) USD/JPY

Figure 5. Results of impulse response function. Notes: Figure 5 shows the results on the response of volatility in spot exchange rates for EUR/USD (**a**) and USD/JPY (**b**) to a one standard deviation shock in unexpected trading volume and unexpected open interest. The response of spot volatility to its own shock is positive and high. It diminishes quickly in the case of EUR/USD, while it takes longer time to die out in the case of USD/JPY.

5. Conclusions

Trading of FX futures in Thailand began in 2012, with only one futures contract available for trading, USD futures. Since 31 October 2022 onwards, Thailand Futures Exchange (TFEX) has added two new FX futures, EUR/USD futures and USD/JPY futures. As concerns about whether FX futures trading has a stabilizing or destabilizing effect on spot volatility, this paper investigates how and to what extent the introduction of FX futures (EUR/USD and USD/JPY futures) has an impact on spot volatility in the context of Thailand and considers the contemporaneous and dynamic relationship between spot volatility and FX futures trading activity, including trading volume and open interest.

The GARCH family models augmented with the dummy variable for the impact of the introduction of FX futures are applied for modelling spot volatility over the sample

period from 27 September 2021 to 12 January 2024. The EGARCH (1,1) model is found to be the best fitted model for both EUR/USD and USD/JPY returns. The results suggest that the introduction of EUR/USD futures and USD/JPY futures by TFEX decreases spot volatility. By dividing the whole sample into two sub-periods (pre and post introduction periods) and applying the EGARCH (1,1) model for the pre and post introduction periods, the results suggest that the introduction of FX futures by TFEX increases the rate at which new information is incorporated into underlying spot prices and decreases the persistency of volatility shocks. This implies an improvement in spot market efficiency as a result of the introduction of FX futures by TFEX. Over the post introduction period from 31 October 2022 to 12 January 2024, trading volume and open interest in EUR/USD and USD/JPY futures are decomposed of expected and unexpected components using the ARIMA model. They are added into a conditional variance equation of the EGARCH (1,1) model for analysis of the contemporaneous relationship between spot volatility and FX futures trading activity. The results show a positive effect of unexpected trading volume and a negative effect of unexpected open interest on contemporaneous spot volatility. These results are in line with the VAR(1) model results of the dynamic relationship between spot volatility and FX futures trading activity. The Granger causality test results also show a significant causality running from unexpected trading volume to spot volatility or unexpected open interest to spot volatility. The results of the impulse response function provide convincing proof of the sign results of the VAR(1) model, which indicate that unexpected trading volume has a destabilizing impact, while unexpected open interest has a stabilizing impact on spot volatility. However, with the impact on spot volatility caused by unexpected open interest being stronger than that by unexpected trading volume, FX futures trading stabilizes spot volatility.

Since the introduction of FX futures can improve spot market efficiency and lower spot market volatility, TFEX should add new FX futures so that investors can select a mix of FX futures that align with their investment objectives and risk tolerance. In addition, it is important to encourage hedgers or informed traders into the FX futures market to ensure a stabilizing impact of FX futures trading on spot volatility. Supporting traders through education/training will enhance their confidence in the use of FX futures. Although this study focuses on the impact of FX futures trading on spot volatility, an interesting extension would be to assess the linkage between the futures market volatility and spot market volatility. In addition, a longer time series would be needed to conduct an analysis.

Funding: This research was funded by Department of Economics, Faculty of Economics, Kasetsart University.

Data Availability Statement: The original data presented in the study are openly available in FigShare at https://doi.org/10.6084/m9.figshare.25912345 (accessed on 28 May 2024).

Conflicts of Interest: The author declares no conflicts of interest.

References

Bessembinder, Hendrik, and Paul J. Sequin. 1992. Futures-Trading Activity and Stock Price Volatility. *Journal of Finance* 47: 2015–34. [CrossRef]

Bhargava, Vivek, and D. K. Malhotra. 2007. The Relationship between Futures Trading Activity and Exchange Rate Volatility, Revisited. *Journal of Multinational Financial Management* 17: 95–111. [CrossRef]

Bollerslev, Tim. 1986. Generalized Autoregressive Conditional Heteroskedasticity. *Journal of Econometrics* 31: 307–27. [CrossRef]

Box, George E. P., and Gwilym M. Jenkins. 1976. *Time Series Analysis: Forecasting and Control*, revised ed. San Francisco: Holden Day.

Chatrath, Arjun, Sanjay Ramchander, and Frank Song. 1996. The Role of Futures Trading Activity in Exchange Rate Volatility. *Journal of Futures Markets* 16: 561–84. [CrossRef]

Conlon, Thomas, Shaen Corbet, and Richard J. McGee. 2024. The Bitcoin Volume-Volatility Relationship: A High Frequency Analysis of Futures and Spot Exchanges. *International Review of Financial Analysis* 91: 103013. [CrossRef]

Fleming, Jeff, and Barbara Ostdiek. 1999. The Impact of Energy Derivatives on the Crude Oil Market. *Energy Economics* 21: 135–67. [CrossRef]

Futures Industry Association. 2024. ETD Tracker. Available online: https://www.fia.org/fia/etd-tracker (accessed on 4 May 2024).

Glosten, Lawrence R., Ravi Jagannathan, and David E. Runkle. 1993. On the Relation between the Expected Value and the Volatility of the Nominal Excess Return on Stocks. *Journal of Finance* 48: 1779–801. [CrossRef]

Gokcan, Suleyman. 2000. Forecasting Volatility of Emerging Stock Markets: Linear versus Non-Linear GARCH Models. *Journal of Forecasting* 19: 499–504. [CrossRef]

Gupta, Ritu. 2017. EURO/INR Futures and Exchange Rate Volatility. *International Journal of Academic Research and Development* 2: 268–70. Available online: https://www.multidisciplinaryjournal.in/assets/archives/2017/vol2issue4/2-4-80-253.pdf (accessed on 1 May 2024).

Guru, Anuradha. 2010. Interplay between Exchange Traded Currency Futures Markets, Spot Markets and Forward Markets: A Study on India. *Indian Economic Review* 45: 111–30.

Jin, Zhongxia, Yue Zhao, and Haobin Wang. 2021. Establishing a Foreign Exchange Futures Market in China. IMF Working Paper No. 2021/268. Available online: https://www.imf.org/-/media/Files/Publications/WP/2021/English/wpiea2021268-print-pdf.ashx (accessed on 5 May 2024).

Jochum, Christian, and Laura Kodres. 1998. Does the Introduction of Futures on Emerging Market Currencies Destabilize the Underlying Currencies? *IMF Staff Papers* 45: 486–521. Available online: https://www.imf.org/external/pubs/ft/wp/wp9813.pdf (accessed on 5 May 2024). [CrossRef]

Jongadsayakul, Woradee. 2020. The Effect of New Futures Contracts on Gold Futures Price Volatility: Evidence from the Thailand Futures Exchange. *Cogent Economics & Finance* 8: 1–14. [CrossRef]

Jongadsayakul, Woradee. 2023. Impact of Derivative Warrants Introduction on Thailand Stock Market Volatility. *International Journal of Business and Society* 24: 1143–56. [CrossRef]

Jongadsayakul, Woradee. 2024. Impact of Night Trading Sessions on Volatility of USD Futures Market in Thailand. *Journal of International Studies* 17: 9–21. [CrossRef]

Kumar, Brajesh. 2009. Effect of Futures Trading on Spot Market Volatility: Evidence from Indian Commodity Derivatives Markets. *SSRN Electronic Journal*. [CrossRef]

Lee, Chyi L., Simon Stevenson, and Ming-Long Lee. 2014. Futures Trading, Spot Price Volatility and Market Efficiency: Evidence from European Real Estate Securities Futures. *Journal of Real Estate Finance and Economics* 48: 299–322. [CrossRef]

Malhotra, Meenakshi, and Dinesh K. Sharma. 2016. Volatility Dynamics in Oil and Oilseeds Spot and Futures Market in India. *Vikalpa* 41: 132–48. [CrossRef]

Miaha, Mamun, and Azizur Rahmanb. 2016. Modelling Volatility of Daily Stock Returns: Is GARCH(1,1) Enough? *American Scientific Research Journal for Engineering, Technology, and Sciences* 18: 29–39. Available online: https://core.ac.uk/download/pdf/235049858.pdf (accessed on 1 May 2024).

Nelson, Daniel B. 1991. Conditional Heteroskedasticity in Asset Returns: A New Approach. *Econometrica* 59: 347–70. [CrossRef]

Oduncu, Arif. 2011. The Effects of Currency Futures Trading on Turkish Currency Market. *Journal of BRSA Banking and Financial Markets* 5: 97–109. Available online: http://www.bddk.org.tr/Content/docs/bddkDergiTr/dergi_0009_06.pdf (accessed on 4 May 2024).

Rani, Priyanka, Sonia, and Karam P. Narwal. 2022. Impact of Currency Futures Issuance on Foreign Exchange Rate Volatility in India. *Orissa Journal of Commerce* 43: 68–84. [CrossRef]

Röthig, Andreas. 2004. Currency Futures and Currency Crises. Darmstadt Discussion Papers in Economics No. 136. Available online: https://www.econstor.eu/bitstream/10419/22519/1/ddpie_136.pdf (accessed on 4 May 2024).

Sahu, Dhananjay. 2012. Dynamics of Currency Futures Trading and Underlying Exchange rate Volatility in India. *Research Journal of Finance and Accounting* 3: 15–23. Available online: https://api.semanticscholar.org/CorpusID:212521438 (accessed on 9 May 2024).

Sakthivel, P., Krishna R. Chittedi, Daniel Sakyi, and V. Vijay Anand. 2017. The Effect of Currency Futures on Volatility of Spot Exchange Rates: Evidence from India. *International Journal of Economic Research* 14: 427–35. Available online: https://www.researchgate.net/publication/320145569_The_effect_of_currency_futures_on_volatility_of_spot_exchange_rates_Evidence_from_India (accessed on 8 May 2024).

Sharma, Somnath. 2011. An Empirical Analysis of the Relationship between Currency Futures and Exchange Rates Volatility in India. RBI Working Paper Series No. 01. Available online: https://rbidocs.rbi.org.in/rdocs/Publications/PDFs/1_FCP010411.PDF (accessed on 1 May 2024).

Shastri, Kuldeep, Jahangir Sultan, and Kishore Tandon. 1996. The Impact of the Listing of Options in the Foreign Exchange Market. *Journal of International Money and Finance* 15: 37–64. [CrossRef]

Shenbagaraman, Premalata. 2003. Do Futures and Options Trading Increase Stock Market Volatility? NSE Research Initiative Working Paper. Available online: https://nsearchives.nseindia.com/content/research/Paper60.pdf (accessed on 8 May 2024).

Singh, Ashwani, and Govind Patra. 2022. Impact of Currency Futures on Spot Rate Volatility in Indian Foreign Exchange Market. *International Journal of Health Sciences* 6: 7431–48. [CrossRef]

Sivarajadhanavel, P., S. Chandrakumarmangalam, and T. Mohanasundaram. 2016. Accessing the Influence of Currency Futures on Spot Exchange Rate: A Case of India. *Asian Journal of Research in Social Sciences and Humanities* 6: 1148–55. [CrossRef]

Smit, E. V. D. M., and M. W. Louw. 1996. The Relationship between Volatility, Volume and Open Interest: Some Evidence from the South African Futures Market. *South African Journal of Business Management* 27: 113–21. [CrossRef]

Szczygielski, Jan J., and Chimwemwe Chipeta. 2023. Properties of Returns and Variance and the Implications for Time Series Modelling: Evidence from South Africa. *Modern Finance* 1: 35–55. [CrossRef]
Yilgor, Ayse G., and Claurinde L. C. Mebounou. 2016. The Effect of Futures Contracts on the Stock Market Volatility: An Application on Istanbul Stock Exchange. *Journal of Business, Economics and Finance* 5: 307–17. [CrossRef]
Zakoian, Jean-Michel. 1990. Threshold Heteroskedastic Model. *Journal of Economic Dynamics and Control* 18: 931–55. [CrossRef]
Zhang, Chuanhai, Huan Ma, Gideon B. Arkorful, and Zhe Peng. 2021. The Impacts of Futures Introduction on Spot Market Volatility: Evidence from the Bitcoin Market. SSRN Electronic Journal. Available online: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3903735 (accessed on 9 May 2024).

MDPI AG
Grosspeteranlage 5
4052 Basel
Switzerland
Tel.: +41 61 683 77 34

Risks Editorial Office
E-mail: risks@mdpi.com
www.mdpi.com/journal/risks

Academic Open
Access Publishing

mdpi.com

ISBN 978-3-7258-7994-6