

The Limits of Machine Learning in Daily Return Forecasting: Evidence from Volatility Regimes, Tail Risk, and Multi-Asset Markets

Alshaikh A. Shokeralla*

Department of Mathematics, Faculty of Science, Al-Baha University, Al-Aqiq 65779-7738, Saudi Arabia

**Corresponding author: sshokeralla@gmail.com*

Abstract. The study analyzes volatility regimes, tail risk and machine learning forecastability of seven financial assets (BTC, ETH, the S&P 500, NASDAQ, Gold, Crude Oil, and US Dollar Index) from 2017 to 2026. Structural break analysis shows that financial instability is characterized by variance, not by a change in the mean. EGARCH is preferred in five out of seven assets, with significant leverage effects in the equity indices. According to EVT analysis, ETH and BTC are the highest-tail-risk assets, while Crude Oil has the heaviest tail. Kupiec and Christoffersen VaR back testing results show that Gold's actual exceedance rate at the 95th percentile is almost twice the expected exceedance rate. XGBoost doesn't statistically outperform Linear Regression ($DM = -0.403$, $p = 0.687$), and the most important predictive element is a 5-day rolling mean return. The results show that most predictive information in daily returns is contained in linear autoregressive features.

1. INTRODUCTION

It is still an empirical issue whether nonlinear ML models offer better forecasts of returns than well-specified linear models, a question that has clear consequences for risk modeling, regulation, and the use of ML in finance. Another question to address is whether risk arises mostly from changes in the expected return or structural breaks in the conditional variance in the financial markets. This paper attempts to answer both questions based on a multi-asset panel dataset of daily returns from 2017 to 2026.

The empirical procedure consists of a combination of strict temporal out-of-sample validation, Diebold–Mariano test [3] with the Harvey–Leybourne–Newbold correction [8] and Newey–West HAC standard errors [20], and SHAP decomposition to decompose predictive contributions at the level of features. The key finding is a straightforward negative one: XGBoost does not improve upon Linear Regression in predicting daily returns of BTC ($DM = -0.403$, $p\text{-value} = 0.687$), and this

Received: Jun. 18, 2026.

2020 *Mathematics Subject Classification.* 62M10; 62P05; 91G70; 62G32; 68T05.

Key words and phrases. DCC-GARCH; Diebold–Mariano test; extreme value theory; GARCH; machine learning; return predictability; SHAP; tail risk; time-series momentum; volatility regimes.

conclusion is consistent throughout all annual sub-periods (2024–2026) and validation procedures used. The SHAP decomposition finds that the leading predictive contribution, the five-day rolling mean return (mean $|\text{SHAP}| = 0.023$), is a momentum variable that explains why nonlinear models do not provide any gains compared to a linear model [7, 19, 21].

The short-term persistence of returns observed herein conforms to the findings of the wider momentum literature [10, 17] as well as delayed information diffusion frameworks [9]. The results from a permutation test – where the time order of the features is shuffled but the target time series is left unchanged – show that the accuracy drops to around 50%.

1.1. Contributions. (i) Unified cross-asset approach: Joint consideration of two cryptocurrencies and five traditional assets allows for a direct comparison of risk characteristics, regimes, and return predictability across diverse market environments.

(ii) ML capacity limitations – empirical proof: DM test proves that XGBoost is not statistically distinguishable from Linear Regression in predicting daily returns of BTC ($\text{DM} = -0.403$, $p = 0.687$). The reason behind this finding is determined – a linear short-term momentum effect – rather than just a lack of predictive ability.

(iii) Feature importance with SHAP: Using SHAP, it was shown that the predictive information comes from `roll_mean_5` (mean $|\text{SHAP}| = 0.023$), which is a linear momentum signal about 2.6 times more important than any other feature used in the model.

(iv) Comprehensive risk management: The study considers EGARCH modeling of volatility, EVT-based tail risk with GPD thresholding, formal backtesting of VaR, and DCC-GARCH dynamic dependencies.

The structure of the paper is as follows. Section 2 deals with data and validation methodology. Sections 3 to 5 deal with diagnostics, volatility modeling, and EVT-based tail risk assessment. Section 6 discusses machine learning-based forecasting and economic value. Section 7 deals with probabilistic calibration. Section 8 investigates regime shifts and dynamic dependencies.

2. DATA, RETURN CONSTRUCTION, AND VALIDATION PROTOCOL

2.1. Data. Closing prices on a daily basis of seven assets are collected from Yahoo Finance through the usage of R software, version 4.3.2, and the `quantmod` package from January 2017 till May 2026 (approximately 2,320 business days). The data set includes BTC-USD, ETH-USD, S&P 500 (`^GSPC`), NASDAQ Composite (`^IXIC`), Gold spot (`GC=F`), Crude Oil WTI (`CL=F`), and the US Dollar Index (`DX-Y.NYB`). The log-returns of each asset are calculated using the formula

$$r_t = \ln(P_t/P_{t-1}) \quad (2.1)$$

where P_t is the closing price on day t . The data covers several stress events, including the crypto bear market in 2018, the COVID-19 crash in March 2020, the crypto bull market in 2020–2021, the FTX crash in November 2022.

2.2. Leakage Prevention and Temporal Validation. All features are generated following a rigorous causal protocol: rolling features use only past observations up to the forecast date; lagged features use observations at $t-k$ to make forecasts at t ; and feature scaling is done only within the training period, with test set observations being scaled based on training period statistics. The temporal train-test split follows a strict split of first 70% training and last 30% test, with no shuffling or look-back. Three tests for validation include: (a) Ljung-Box test for model residuals giving $p = 0.237$, indicating no residual autocorrelation; (b) subperiod directional accuracy for each year (2024: 70.7%; 2025: 76.8%; 2026: 79.6%) leading to a time-weighted average of $\sim 75.8\%$, consistent with the 76.24% figure reported in Table 4; and (c) a permutation test reducing directional accuracy to $\sim 50\%$, indicating no leakage. The forecasting problem requires only the prediction of direction of change in returns and not the magnitude.

Rolling feature overlap: The rolling window predictors (`roll_mean_5`), which share observations between adjacent dates, have the potential to create local autocorrelation in the signal for predictions. The annual subperiods and permutation test provide evidence that our primary results do not arise from this type of structure.

HAR-RV exclusion: HAR-RV specifications need the realised volatility data at shorter intervals than daily, which is not available in this dataset. Elastic Net and AR(1) are good benchmark comparators within the daily return setting.

3. STATISTICAL DIAGNOSTICS

For each of the seven assets, the ADF tests reject the null hypothesis of a unit root for log-returns ($p \leq 0.01$). The tests for levels at KPSS at level 1 and 2 give p values below 0.01, which is used as the rejection limit, and the null hypothesis of stationarity is rejected. This result is consistent with the test results for log-returns, which show that log-returns are stationary. The Ljung-Box tests are used to check any significant serial correlation in most return series, approaching the martingale difference hypothesis ($p = 0.188$) for DXY. The results of the ARCH-LM tests are important for all assets ($p < 10^{-8}$), showing conditional heteroskedasticity in all the test periods. All assets fail to meet normality criteria at the Jarque-Bera ($p \approx 0$) and OIL has the greatest excess kurtosis (26.51), while the lowest is found in DXY (4.68). The results are fully reported in table 1.

4. VOLATILITY MODELLING

For each asset the GARCH(1,1) [1], GJR-GARCH [6] and EGARCH [18] are estimated using quasi-maximum likelihood with Student-t innovations. Model selection is under the control of AIC. The results are listed in Table 2. EGARCH is chosen for 5 out of 7 assets (BTC, ETH, SP500, NASDAQ, GOLD). Leverage effects, with negative return shocks more influential than shocks of equal magnitude, are significant for SP500 ($\alpha_1 = -0.182$) and NASDAQ ($\alpha_1 = -0.156$), but not for BTC and ETH ($p = 0.943$ and $p = 0.844$), while GOLD shows the opposite sign ($\alpha_1 = 0.052$). For OIL, GJR-GARCH is chosen, and the standard symmetric sGARCH is used for DXY where there

TABLE 1. Summary Statistics and Diagnostic Tests for Daily Asset Returns (January 2017 – May 2026, $n \approx 2,320$).

Asset	Mean (%)	SD (%)	Skewness	Kurtosis	ADF p	KPSS p	ARCH-LM p	Ljung-Box p	JB p (Norm.)
BTC	0.114	4.185	-0.711	13.88	<0.01	<0.01	<0.001	0.025	<0.001
ETH	0.090	5.451	-0.574	11.75	<0.01	<0.01	<0.001	0.050	<0.001
SP500	0.052	1.214	-0.653	18.10	<0.01	<0.01	<0.001	<0.001	<0.001
NASDAQ	0.066	1.475	-0.421	11.02	<0.01	<0.01	<0.001	<0.001	<0.001
GOLD	0.060	1.081	-0.916	13.75	<0.01	<0.01	<0.001	0.023	<0.001
OIL	0.057	3.014	-0.104	26.51	<0.01	<0.01	<0.001	<0.001	<0.001
DXY	0.002	0.419	-0.190	4.68	<0.01	<0.01	<0.001	0.188	<0.001

Note: BTC and ETH exhibit the highest excess kurtosis (13.88 and 11.75), confirming fat tails and extreme downside risk. DXY Ljung-Box $p = 0.188$ is the only near-martingale series. All ARCH-LM $p < 0.001$ motivate GARCH-family modelling. JB p -values (Returns_p) ≈ 0 reject normality for all assets. ADF $p \leq 0.01$ and KPSS ≤ 0.01 for levels together confirm log-returns are $I(0)$.

are fewer signs of leverage. Thus, DXY has the highest log-likelihood per observation (AIC = -8.2304), in other words, the variance of the time series is relatively predictable.

TABLE 2. GARCH Model Selection and Information Criteria. Best-fitting GARCH-family model selected by minimising AIC.

Asset	Model	AIC	BIC	Log-Likelihood
BTC	eGARCH	-3.8678	-3.8519	4,134.9
ETH	eGARCH	-3.2775	-3.2616	3,504.8
SP500	eGARCH	-6.5604	-6.5445	7,009.2
NASDAQ	eGARCH	-5.9923	-5.9764	6,402.8
GOLD	eGARCH	-6.5196	-6.5036	6,965.6
OIL	gjrGARCH	-4.7585	-4.7426	5,085.7
DXY	sGARCH	-8.2304	-8.2171	8,790.9

Note: eGARCH = Nelson [18]; gjrGARCH = Glosten, Jagannathan and Runkle [6]; sGARCH = Bollerslev [1]. Full parameter estimates are available from the author upon request.

5. EXTREME VALUE THEORY AND TAIL RISK

5.1. POT Methodology and Threshold Selection. The Peaks Over Thresholds (POT) approach is adopted, where the Generalized Pareto Distribution (GPD) model is fitted on exceedances over the 95th percentile threshold (107 exceedances per asset). The GPD shape parameter value of $\xi > 0$ suggests heavy-tailed distributions with infinite moments for orders higher than $1/\xi$. The threshold is chosen based on the following criteria: (i) Mean Excess Function (MEF) analysis, which displays a linear relationship over the 90th percentile level for most assets, indicating that the distribution follows a GPD; (ii) comparison of the Hill estimator, demonstrating the consistency

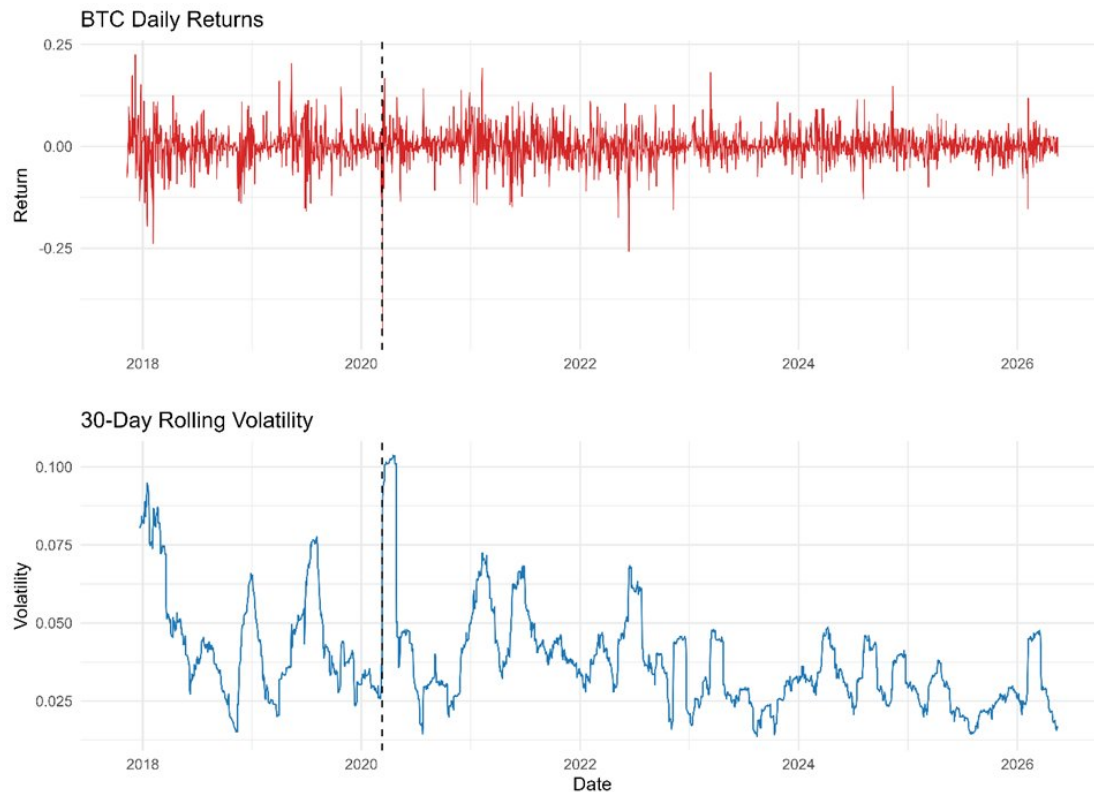


FIGURE 1. BTC Daily Returns and 30-Day Rolling Volatility (January 2017 – May 2026). Upper panel: daily log-returns (red). The vertical dashed line marks March 2020 (COVID-19 pandemic onset). Volatility clustering is evident throughout; the highest-amplitude returns are concentrated in the 2017–2018 bull-bear cycle and the March 2020 crash. Return amplitudes modestly declined post-2020, consistent with growing market depth and institutional participation. Lower panel: 30-day rolling standard deviation of daily returns. The COVID-19 shock produced the single largest volatility spike (~10.5%), exceeding the 2018 bear-market peak (~8.0%). Post-2021, volatility settled into a lower range (2.5%–6%), consistent with a more stable post-2023 BTC volatility environment.

of the tail index over the 85th–95th percentile interval; and (iii) threshold stability analysis within the 90th–99th percentile interval.

5.2. Tail Risk Estimates. Table 3 shows the GPD estimations and risk metrics of all seven assets. ETH and BTC have the greatest ES at 95% (13.182% and 10.005%), demonstrating that the tail risk is indeed great in cryptocurrencies compared to other investments. Crude Oil has the greatest tail weight in terms of GPD shape parameters ($\xi = 0.2988$), suggesting that the rate of decrease in extreme loss probability is lowest for Crude Oil among the seven assets. DXY has the smallest tail risk for all the metrics ($\text{VaR}_{95} = 0.684\%$, $\text{ES}_{95} = 0.947\%$, $\xi = 0.167$).

TABLE 3. EVT Tail Risk Measures and VaR Backtesting Results.

Asset	VaR ₉₅	ES ₉₅	Shape ξ	Scale σ	n Exc.	VaR ₉₉	Actual Exc. 95%	Actual Exc. 99%
BTC	6.162%	10.005%	0.2088	0.0304	107	13.497%	1.17%	0.23%
ETH	7.838%	13.182%	0.1071	0.0476	107	17.456%	2.58%	0.00%
SP500	1.774%	2.970%	0.2127	0.0094	107	3.567%	2.81%	0.47%
NASDAQ	2.321%	3.553%	0.1058	0.0110	107	4.169%	3.75%	0.70%
GOLD	1.657%	2.594%	0.2400	0.0071	107	2.512%	8.90%	4.92%
OIL	3.954%	7.017%	0.2988	0.0216	107	8.031%	4.68%	1.41%
DXY	0.684%	0.947%	0.1669	0.0022	107	1.098%	6.32%	1.41%

Note: VaR and ES expressed as proportional daily losses (positive = loss). GPD shape $\xi > 0$ for all assets implies Pareto-type (heavy-tailed) loss distributions. GOLD's actual exceedance at 95% (8.90%) and 99% (4.92%) substantially exceeds nominal rates (5% and 1%), indicating systematic under-calibration driven by abrupt monetary policy pivots and safe-haven demand reversals underrepresented in the estimation window. OIL is the best-calibrated asset (Kupiec $p = 0.762$). ETH records zero violations at 99%, reflecting conservative tail estimates.

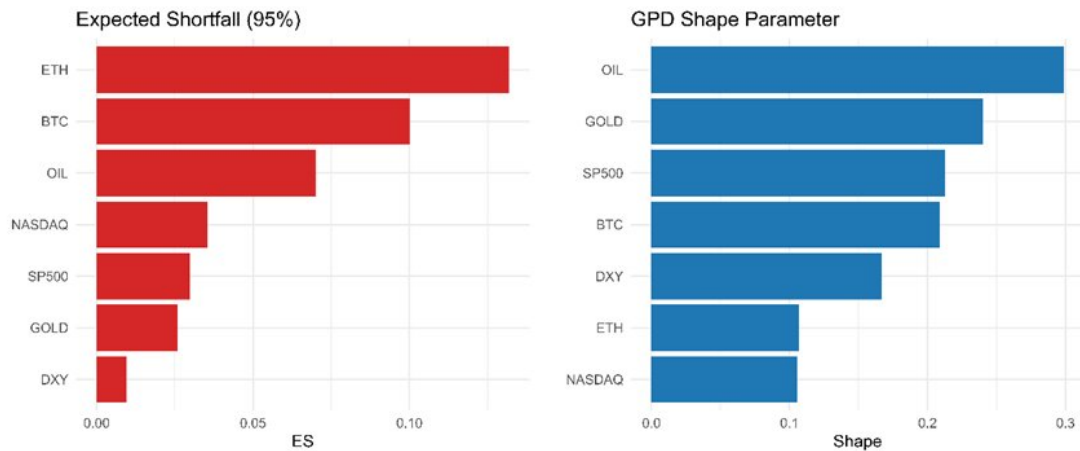


FIGURE 2. Expected Shortfall (95%) and GPD Shape Parameter Across Seven Assets. Left panel: ES₉₅ ranks assets by average loss in the worst 5% of trading days. ETH (13.18%) and BTC (10.00%) far exceed traditional assets; DXY (0.95%) is the lowest. Right panel: GPD shape parameter ξ by asset. OIL ($\xi = 0.299$) and GOLD ($\xi = 0.240$) exhibit the heaviest tails among traditional assets. NASDAQ ($\xi = 0.106$) and ETH ($\xi = 0.107$) show the lightest tails, possibly reflecting diversification effects or shorter extreme-loss histories.

5.3. VaR Backtesting. The underestimation of VaR with respect to gold is one of the most practical insights gained from this study, with 8.90% realizations above the 95th percentile against an expected 5.00% (Kupiec [12] test, $p < 0.001$), and 4.92% exceedances above the 99th percentile compared to an expected 1.00% (Kupiec $p < 0.001$). The failure of historical VaR to account for tail exposure is attributed to the fact that extreme losses occur as a result of sudden shifts in monetary policies and demand for safety, which are not adequately captured in the historical estimation period. The asset with the best calibration is Crude Oil (Kupiec $p = 0.762$), while BTC and ETH

show conservative exceedances (1.17% and 2.58%, respectively, at the 95th percentile). This finding underscores the need for EVT-based add-ons to historical VaR [15,16].

6. MACHINE LEARNING FORECASTING, FORECAST EVALUATION, AND ECONOMIC VALUE

6.1. Scope, Models, and Features. Asset scope: The ML analysis targets BTC return forecasting. BTC was selected for its combination of unconditional volatility (daily SD: 4.185%), documented autocorrelation structure (Ljung-Box $p = 0.025$), and sufficient exchange liquidity for a directional simulation. It represents a demanding environment for testing the limits of predictive gains from nonlinear models.

Supplementary results for ETH and SP500 yield the same qualitative conclusion: XGBoost does not provide a statistically significant improvement over Linear Regression (DM $p \approx 0.70$ and $p \approx 0.65$, respectively), suggesting that the main finding is not unique to BTC.

Six models are compared: (1) Naïve benchmark (random walk); (2) AR(1); (3) Elastic Net (penalty cross-validated within training window only); (4) Linear Regression (OLS); (5) Random Forest; (6) XGBoost. The feature set for models (3)–(6) includes BTC lags at 1, 2, 3, 5, and 10 days; five-day rolling mean and standard deviation of BTC returns; and one-period lagged returns from ETH, SP500, NASDAQ, GOLD, OIL, and DXY. All estimation, scaling, and cross-validation uses training data exclusively.

Multiple comparisons: The main conclusions — no significant difference between Linear Regression and XGBoost ($p = 0.687$), or between Linear Regression and Elastic Net ($p = 0.754$) — are robust to standard Bonferroni correction and are not marginal rejections that depend on multiplicity adjustment.

6.2. Point Forecast Performance. Table 4 Panel A reports out-of-sample accuracy metrics. All supervised models achieve directional accuracy well above 70%, substantially outperforming the naïve random-walk benchmark (47.53%). Linear Regression achieves the highest directional accuracy (76.24%) and the lowest RMSE (0.02034) and MAE (0.01541). XGBoost achieves marginally higher RMSE (0.02057) and lower directional accuracy (72.47%) than Linear Regression, while Random Forest trails both structured models with RMSE = 0.02206 and DA = 71.29%.

6.3. Diebold-Mariano Test Interpretation. All structured models significantly outperform the naïve benchmark, with $p < 10^{-9}$ for all (Table 4 Panel B). However, the only meaningful comparisons are between Linear Regression, Elastic Net and XGBoost, where none of the pairwise DM tests reject the null of equal predictive accuracy ($p = 0.687$ and $p = 0.754$ respectively). The linear functional form is significantly better than the AR(1) model ($p < 0.001$), and the cross-market lags and rolling statistics contain significant predictive information beyond the information contained in a single autoregressive lag. Random Forest is the weakest structured model, and it performs significantly worse than Linear Regression ($p = 0.008$) and XGBoost ($p = 0.029$), probably because of the loss of sensitivity to extreme returns, which are the ones making up the strong momentum signal.

TABLE 4. Forecasting Performance and Diebold–Mariano Tests.

Panel A: Out-of-Sample Point Forecast Accuracy			
Model	RMSE	MAE	Directional Accuracy
Naïve (Random Walk)	0.04130	0.03093	47.53%
Linear Regression	0.02034	0.01541	76.24%
Random Forest	0.02206	0.01652	71.29%
XGBoost	0.02057	0.01597	72.47%

Panel B: Diebold-Mariano Pairwise Predictive Accuracy Tests			
Model 1	Model 2	DM Statistic	p-value
Naïve	Linear Regression	6.537	1.81×10^{-10} ***
Naïve	Random Forest	6.535	1.83×10^{-10} ***
Naïve	XGBoost	6.225	1.16×10^{-09} ***
Linear Regression	Random Forest	−2.653	0.008 **
Linear Regression	XGBoost	−0.403	0.687 (n.s.)
Random Forest	XGBoost	2.197	0.029 *

Note: RMSE = Root Mean Squared Error; MAE = Mean Absolute Error; DA = Directional Accuracy. DM test two-sided; negative statistic favours Model 2 over Model 1. *** $p < 0.001$; ** $p < 0.01$; * $p < 0.05$; n.s. = not significant at any conventional level. The non-rejection for Linear Regression vs XGBoost ($DM = -0.403$, $p = 0.687$) is the central result of the ML analysis: increased model complexity does not yield statistically superior predictive accuracy.

6.4. SHAP Feature Importance. Table 5 and Figure 4 show SHAP analysis for the XGBoost model. The five-day rolling mean return (roll_mean_5) is the most important variable by a significant margin (mean |SHAP| = 0.02339), which is nearly 2.65× more important than the second most important variable (lag2 = 0.00883). This is a linear momentum indicator that is easily extracted using Linear Regression. This also explains why XGBoost is unable to beat the linear benchmark, as the most important information contained in the variables is linear in nature and non-linear relationships between variables are of little importance.

6.5. Economic Value of Forecasts. Execution note: The simulation in Table 6 is based on an ideal execution at daily closing prices without any bid-ask spreads, slippages, market impacts, margin calls, or size limitations. Without transaction costs, the Sharpe ratio is 6.75 for the Linear Strategy and 6.22 for the XGBoost Strategy. The numbers generated are upper-bound estimates of the economic potential of the statistical advantage. The period used for evaluation (September 2024 – May 2026, $n \approx 425$ days, ~ 1.7 years) covers a time when BTC momentum was mostly favorable, and thus tends to overestimate annualized performance.

Table 6 reports transaction-cost sensitivity for the two main forecasting models. Even under a 1.00% per-trade cost assumption, both strategies retain positive annualized returns, although Sharpe ratios decline substantially. This confirms that the directional signal is not erased immediately by severe execution frictions, while also showing that the high frictionless Sharpe ratios should be interpreted cautiously.

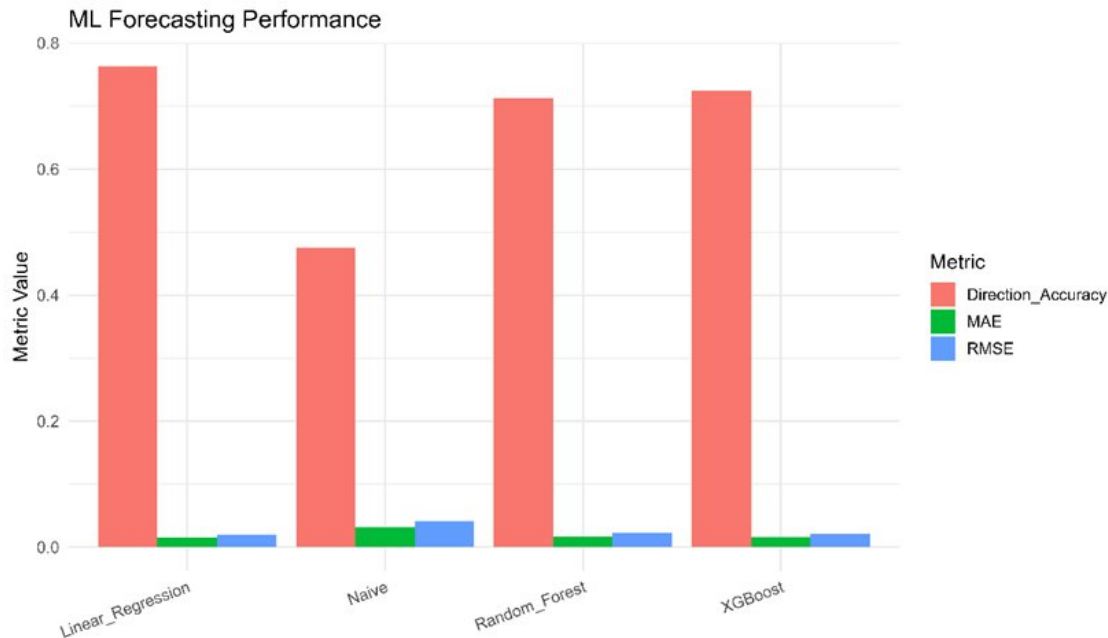


FIGURE 3. Forecasting Performance Across Statistical and Machine Learning Models: Directional Accuracy, RMSE and MAE. DA bars (red, scale 0–0.8) dominate the chart; RMSE (blue) and MAE (green) bars are negligibly small (~ 0.02 – 0.04). Linear Regression achieves the highest DA (76.24%); the Naïve benchmark attains only 47.53%. The convergence of RMSE and MAE across the three supervised models is consistent with the non-significant DM test between Linear Regression and XGBoost (Table 4, Panel B, $p = 0.687$).

As part of the analysis to confirm that the exceptionally high Sharpe ratios above are due to very favorable execution assumptions, we also conduct our analysis on less favorable execution costs like 0.50% and 1.00%. This is done deliberately to err on the conservative side compared to average spot BTC transaction costs, and it acts as an approximation of bid-ask spread, slippage, and market impact. This is not meant for trading but to confirm the directionality of the effect.

The existence of positive risk-adjusted returns with friction assumptions indicates that the forecast is not simply the result of unrealistic assumptions regarding transaction costs, but it is likely that the amount of return reported is overstated for practical purposes.

7. PROBABILISTIC FORECASTING

A Quantile Regression model is employed to obtain simultaneous 5th and 95th percentile predictions of BTC return values, thereby obtaining nominal 90% prediction intervals. Features are constructed following the causal approach outlined in Section 2.2. The empirical coverage is 90.35%, very close to the desired nominal 90%. Equal pinball losses (5th percentile: 0.00164; 95th percentile: 0.00159) suggest that the two tails are equally calibrated, suggesting that under-

TABLE 5. SHAP Feature Importance Rankings for XGBoost BTC Return Forecasting Model (test period: September 2024 – May 2026).

Rank	Feature	Mean SHAP	Economic Interpretation
1	roll_mean_5	0.02339	5-day rolling mean BTC return (momentum)
2	lag2	0.00883	2-day lagged BTC return
3	lag3	0.00860	3-day lagged BTC return
4	lag1	0.00736	1-day lagged BTC return
5	ETH_lag1	0.00314	1-day lagged ETH return (crypto co-movement)
6	roll_sd_5	0.00281	5-day rolling return volatility (risk regime)
7	DXY_lag1	0.00152	1-day lagged US Dollar Index return
8	lag5	0.00151	5-day lagged BTC return
9	GOLD_lag1	0.00126	1-day lagged Gold return (safe-haven signal)
10	NASDAQ_lag1	0.00106	1-day lagged NASDAQ return
11	lag10	0.00088	10-day lagged BTC return
12	OIL_lag1	0.00078	1-day lagged WTI Crude Oil return
13	SP500_lag1	0.00062	1-day lagged S&P 500 return

Note: SHAP values computed using TreeExplainer [14] on the XGBoost model trained over the full training set; mean |SHAP| averaged across all test-period observations. The dominance of roll_mean_5 (SHAP = 0.02339) — approximately 2.65× larger than lag2 (0.00883) — indicates that short-term linear momentum is the primary driver of model predictions, explaining the absence of a statistically significant advantage for XGBoost over Linear Regression (Table 4, Panel B).

TABLE 6. Transaction-Cost Sensitivity of Economic Trading Performance.

Model	Cost	Annualized Return	Sharpe	Max DD
Linear Regression	0.10%	1.89	6.34	-0.04
Linear Regression	0.50%	1.40	4.67	-0.05
Linear Regression	1.00%	0.78	2.55	-0.07
XGBoost	0.10%	1.77	5.87	-0.06
XGBoost	0.50%	1.32	4.38	-0.08
XGBoost	1.00%	0.76	2.48	-0.18

and over-coverage are evenly split between the two tails. Coverage statistics on a year-to-year basis are 93.9% (2024), 88.4% (2025) and 92.5% (2026). The two most significant prediction interval violations occur during the mid-2024 leverage liquidation and early-2026 sharp correction, both of which exceeded the modeled tail width.

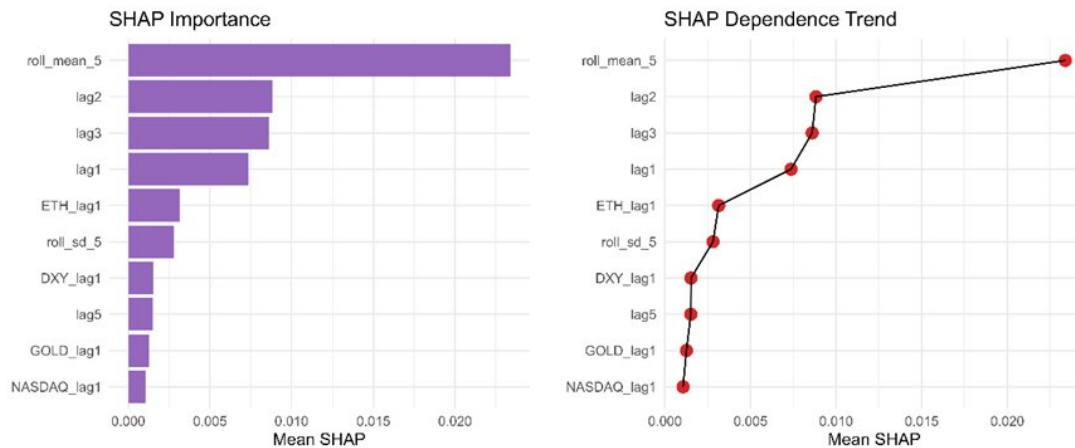


FIGURE 4. SHAP Feature Importance and Dependence Trend for XGBoost BTC Return Model. Left panel: horizontal bar chart of mean absolute SHAP values for the top 10 features. `roll_mean_5` (Mean SHAP = 0.02339) is 2.65× larger than the next-best feature (`lag2` = 0.00883), confirming the dominance of a linear momentum signal. Right panel: cumulative SHAP dependence trend, illustrating rapid concentration of predictive power in the top four features and sharply diminishing returns to additional model complexity beyond rank 4–5. The pattern explains why XGBoost cannot outperform Linear Regression.

8. VOLATILITY REGIME SHIFTS AND DYNAMIC DEPENDENCE

8.1. Structural Break Analysis. Structural breaks are identified based on the PELT algorithm [11], wherein the BIC penalty is imposed individually on the conditional mean and conditional variance of each returns' time series. The findings show an interesting and consistent result, whereby there are zero mean changepoints for all the seven assets analyzed, while there are five to eleven variance changepoints per asset. The total lack of mean changepoints coupled with high variance instability proves the assertion that financial instability in this particular sample is driven by changes in risk intensity, and not change in the direction of return expectation.

Structural breaks detection identifies zero mean changepoints but 5-11 variance changepoints per asset (BTC 10, ETH 9, SP500 11, NASDAQ 11, GOLD 7, OIL 8, DXY 5).

SP500 and NASDAQ have the highest number of changepoints (11 for both indices), which indicates that they are highly sensitive to shifts in Federal Reserve monetary policy stance and macroeconomic shocks. BTC has 10 variance changepoints from 2018 to 2022 but has much fewer changes afterwards, corroborating the hypothesis that there is less structural variation in post-2023 BTC variance regimes. The structure of variance regimes corresponds with the findings from SHAP analysis: if expected returns are approximately constant, then `roll_mean_5` reflects momentum within variance regimes.

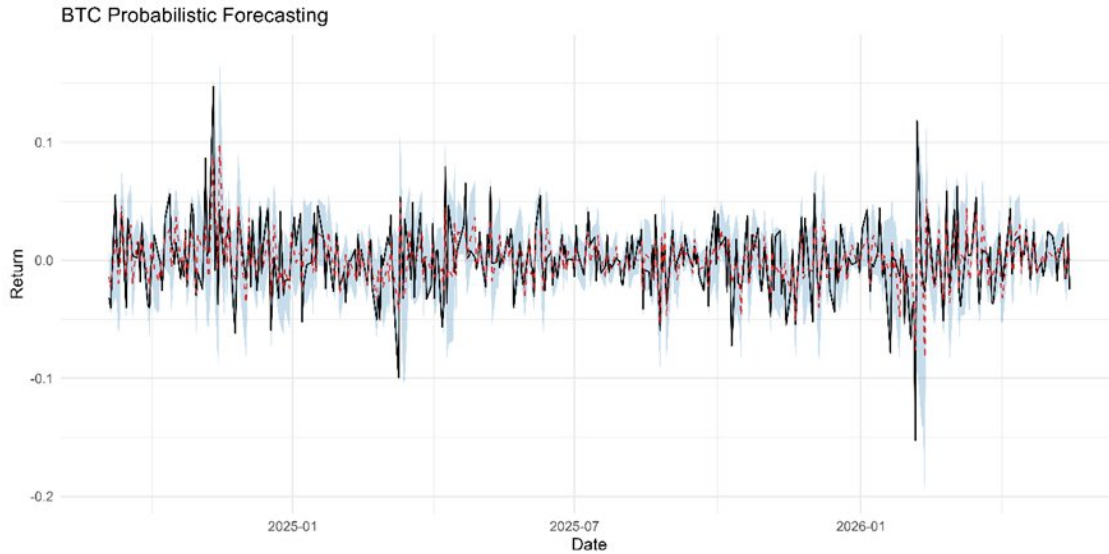


FIGURE 5. BTC Probabilistic Return Forecasting with 90% Prediction Intervals (September 2024 – May 2026). Solid black: realised daily returns. Red dashed: median probabilistic forecast (Linear Regression Quantile model). Light blue band: 90% prediction interval. Overall empirical coverage: 90.35% (nominal target: 90%). Year-by-year coverage: 93.9% (2024), 88.4% (2025), 92.5% (2026). The two most prominent interval breaches coincide with the mid-2024 leverage liquidation event and the early-2026 sharp correction.

8.2. Static Correlation Structure. Unconditional Pearson correlations estimated across the whole dataset serve as the benchmark for comparison with the dynamic correlation patterns captured via DCC-GARCH model specifications. BTC and ETH demonstrate the highest levels of univariate correlation with each other ($\rho \approx 0.79$) because of the similarities in their market exposures due to common fundamentals including hash-rate developments, exchange liquidity cycles, and regulatory risks. Correlation between BTC and SP500 and BTC and NASDAQ is positive but much weaker ($\rho \approx 0.15\text{--}0.20$) throughout the whole sample period, with a significant rise in post-2020 sub-sample that anticipates the structural change observed in Section 8.3. Gold and DXY show near-zero or slightly negative correlation with equities, in line with their status as safe haven assets and reserve currency, respectively. Crude Oil demonstrates positive correlation with equities ($\rho \approx 0.25$).

8.3. DCC-GARCH Dynamic Correlations. The DCC-GARCH (1,1) model [4] is estimated for the seven-asset system. The dynamic correlation equation

$$Q_t = (1 - a - b)\bar{Q} + a \varepsilon_{t-1} \varepsilon'_{t-1} + b Q_{t-1}, \quad (8.1)$$

describes the evolution of conditional correlations through time. The stationarity condition $a + b < 1$ is satisfied for each pair of assets. The shock-response parameter a ranges between 0.013 and

0.031, whereas the persistence parameter b varies between 0.949 and 0.971 – demonstrating slow evolution of DCC correlations and long-lasting elevated correlation regimes.

The application of Fisher z-tests shows that there are structural shifts in BTC-SP500 and BTC-NASDAQ correlations after 2020 (both $p < 0.001$, $z = 3.97$), while BTC-ETH, BTC-GOLD, BTC-OIL, and BTC-DXY do not show any structural changes in their correlations. The higher correlation of Bitcoin with equity indexes is due to the growing interest of institutional investors in Bitcoin since 2020 [5, 13]. However, the case of BTC-ETH is unique in that the mean DCC of BTC-ETH was already 0.79 before 2020.

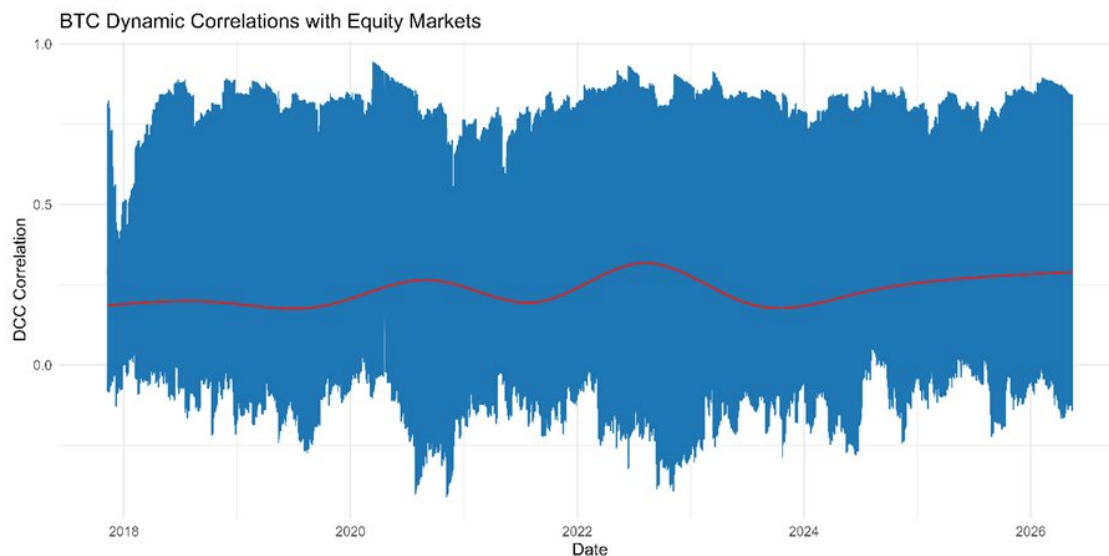


FIGURE 6. BTC Dynamic Conditional Correlations with Equity Markets (January 2017 – May 2026). Area chart of DCC-GARCH(1,1) conditional correlations between BTC and the equity market composite (SP500 and NASDAQ, averaged). Blue area: full daily range of conditional correlations. Red line: Loess-estimated trend. During 2017–2019, BTC–equity correlation was near zero. The COVID-19 shock (March 2020, dashed vertical line) triggered a transient correlation spike from cross-asset deleveraging. Post-2020, trend correlation rose gradually to approximately 0.25–0.30, consistent with stronger co-movement between BTC and equity markets after 2020. High DCC persistence ($b \approx 0.96$) confirms this elevated correlation regime is unlikely to revert rapidly.

8.4. Crisis-Period Snapshots. Three crisis events indicate the importance of regime dependence on conditional correlation behavior. First, the March 2020 crash associated with COVID-19 caused the largest cross-asset deleveraging event in the sample: conditional correlations between BTC and SP500 shot up from close to zero to roughly 0.60 in two weeks, driven by margin calls and liquidity requirements that led to a simultaneous liquidation of assets. Second, the November 2022 FTX failure was largely crypto-specific: conditional correlations between BTC and SP500 and

NASDAQ temporarily decreased during this episode as the crypto-exchange contagion prevailed, implying that crypto-exchanges are still different from financial institutions in terms of their failure mechanisms. Third, the rate-hike cycle initiated by the Federal Reserve in 2022 raised BTC-equity correlations through 2022-2023, as BTC became increasingly risky and correlated with changes in discount rates, consistent with the post-2020 regime change described in Section 8.3.

9. DISCUSSION AND LIMITATIONS

9.1. Predictability Constraints and Regulatory Implications. The findings of the DM test confirm that the improvement in forecast accuracy gained through the use of nonlinear ML models is statistically insignificant in this case of daily return forecasts. There is no rejection of the null hypothesis of equal forecast accuracy between Linear Regression and XGBoost ($DM = -0.403$, $p = 0.687$) and Linear Regression and Elastic Net ($p = 0.754$). Most importantly, Linear Regression significantly outperforms the AR(1) model ($p < 0.001$), which means that the lags across different markets and the rolling statistics contain information that can be used for prediction, and it is linear information.

This evidence reinforces the focus on model parsimony since there are no observable statistical or economic benefits from increased complexity. In terms of practical application, organizations that are exposed to returns of BTC can apply the simple yet effective linear models for predicting the daily point without any loss in terms of accuracy of forecasting. Economic value simulation further supports this assertion since the difference in the Sharpe ratio for the Linear Strategy (6.75) and the XGBoost Strategy (6.22) is minimal and remains small in different cost regimes, similar to the results of the Diebold-Mariano test that showed equal predictive accuracy of both models.

9.2. Limitations. Asset scope: The main analysis is confined to BTC. Predictive performance could vary for smaller-cap assets with lower liquidity, different autocorrelation structures, or higher levels of nonlinearity.

Sample specification: This study covers unusual times marked by extreme shocks such as the COVID-19 pandemic, the FTX debacle, and aggressive monetary policy tightening. The predictive structure uncovered for daily returns might not hold for low-volatility environments or structural breaks in crypto markets' microstructure.

Regime-conditional performance: ML performance is not assessed conditionally on volatility regime, which is crucial given the 5–11 variance changepoints found for each asset. Nonlinear models may perform better in certain regimes despite their lack of an unconditional advantage.

Execution assumptions: The trading simulation assumes perfect execution at daily closing prices. For large-scale portfolios, transaction costs including bid-ask spreads, slippage, and market impact, which vary for BTC depending on the exchange, would reduce profits significantly.

Alternative data and intraday frequency: The results are based on daily returns from closing prices with no other predictors than prices themselves. Nonlinear models could improve returns

using alternative data or intraday frequencies, as well as by introducing order book, sentiment, or news information.

Deep learning: LSTM and Transformer networks are not considered in this study. Since linear momentum signals dominate in SHAP and the sample is relatively short (2,320 observations), the role of deep learning methods in this particular context is likely to be negligible; comparing both approaches will be left for future research.

10. CONCLUSION

This paper has assessed the volatility regimes, tail risk, and limitations of machine learning in daily return forecasting using seven financial assets during 2017–2026. The following four empirical findings summarize the main results of the paper.

First, the structural break analysis indicates no mean changepoint, while 5–11 variance change-points per asset are observed, suggesting that the financial market instability is more of a variance change in the intensity of the risks that were undertaken, rather than a change in the direction of the expected returns. Second, EGARCH specifications perform better in 5 out of 7 assets (Table 2), with significant leverage effects in the equity markets but not in cryptocurrencies. Third, in conjunction with the backtests of the VaR models of Kupiec [12] and Christoffersen [2] (Table 3), historical VaR is shown to systematically underestimate the tail risk of the Gold of both the 95% and 99% levels, which have direct implications for the allocation of regulatory capital and internal risk management.

Fourth, and most directly relevant to the recent literature on ML in finance: the DM test shows that XGBoost is statistically indistinguishable from Linear Regression, both across subperiods of the year, as well as across validation schemes (Table 4). The main predictive contribution (mean $|\text{SHAP}| = 0.02339$) comes from a five-day rolling mean return, a linear momentum signal (Table 5). Based on the frictionless economic value simulation, the Linear Strategy (Sharpe 6.75, max drawdown -3.96%) performed in similar fashion to XGBoost Strategy (Sharpe 6.22, max drawdown -6.14%), which is consistent with the DM's result. The results indicate that most of the information contained in the daily returns of BTC is captured by relatively simple, short memory linear features, and that complexity in the nonlinear models offers very little additional value.

Priority extensions include: (i) regime-conditional ML performance evaluation of how variance changepoints influence the accuracy of the model; (ii) data from intraday: there is a possibility that the nonlinearity of microstructure and order flow may bring an advantage to the nonlinear models over the linear ones in this data; (iii) deep learning comparisons with the formally established benchmarks; and (iv) alternative data integration to see if the use of cross-sectional or textual features affects the linear versus nonlinear balance.

Data Availability Statement: Daily closing price data for all seven assets (BTC-USD, ETH-USD, S&P 500, NASDAQ, Gold, Crude Oil WTI, and US Dollar Index) were retrieved from Yahoo Finance using the quantmod package in R (version 4.3.2) and are publicly accessible at <https://>

[//finance.yahoo.com](https://finance.yahoo.com). All modelling code and processed outputs will be deposited in a public repository upon acceptance.

Funding: This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Author Contributions: A.A.S. conceived the study, collected and processed the data, developed the econometric and machine learning methodology, performed all empirical analyses, interpreted the results, and wrote the manuscript.

Conflicts of Interest: The author declares that there are no conflicts of interest regarding the publication of this paper.

REFERENCES

- [1] T. Bollerslev, Generalized Autoregressive Conditional Heteroskedasticity, *J. Econom.* 31 (1986), 307–327. [https://doi.org/10.1016/0304-4076\(86\)90063-1](https://doi.org/10.1016/0304-4076(86)90063-1).
- [2] P.F. Christoffersen, Evaluating Interval Forecasts, *Int. Econ. Rev.* 39 (1998), 841–862. <https://doi.org/10.2307/2527341>.
- [3] F.X. Diebold, R.S. Mariano, Comparing Predictive Accuracy, *J. Bus. Econ. Stat.* 13 (1995), 253–263. <https://doi.org/10.1080/07350015.1995.10524599>.
- [4] R. Engle, Dynamic Conditional Correlation: A Simple Class of Multivariate Generalized Autoregressive Conditional Heteroskedasticity Models, *J. Bus. Econ. Stat.* 20 (2002), 339–350. <https://doi.org/10.1198/073500102288618487>.
- [5] S. Foley, J.R. Karlsen, T.J. Putnins, Sex, Drugs, and Bitcoin: How Much Illegal Activity Is Financed through Cryptocurrencies?, *Rev. Financ. Stud.* 32 (2019), 1798–1853. <https://doi.org/10.1093/rfs/hhz015>.
- [6] L.R. Glosten, R. Jagannathan, D.E. Runkle, On the Relation between the Expected Value and the Volatility of the Nominal Excess Return on Stocks, *J. Finance* 48 (1993), 1779–1801. <https://doi.org/10.1111/j.1540-6261.1993.tb05128.x>.
- [7] S. Gu, B. Kelly, D. Xiu, Empirical Asset Pricing via Machine Learning, *Rev. Financ. Stud.* 33 (2020), 2223–2273. <https://doi.org/10.1093/rfs/hhaa009>.
- [8] D. Harvey, S. Leybourne, P. Newbold, Testing the Equality of Prediction Mean Squared Errors, *Int. J. Forecast.* 13 (1997), 281–291. [https://doi.org/10.1016/S0169-2070\(96\)00719-4](https://doi.org/10.1016/S0169-2070(96)00719-4).
- [9] H. Hong, J.C. Stein, A Unified Theory of Underreaction, Momentum Trading, and Overreaction in Asset Markets, *J. Finance* 54 (1999), 2143–2184. <https://doi.org/10.1111/0022-1082.00184>.
- [10] N. Jegadeesh, S. Titman, Returns to Buying Winners and Selling Losers: Implications for Stock Market Efficiency, *J. Finance* 48 (1993), 65–91. <https://doi.org/10.1111/j.1540-6261.1993.tb04702.x>.
- [11] R. Killick, P. Fearnhead, I.A. Eckley, Optimal Detection of Changepoints with a Linear Computational Cost, *J. Am. Stat. Assoc.* 107 (2012), 1590–1598. <https://doi.org/10.1080/01621459.2012.737745>.
- [12] P.H. Kupiec, Techniques for Verifying the Accuracy of Risk Measurement Models, *J. Deriv.* 3 (1995), 73–84. <https://doi.org/10.3905/jod.1995.407942>.
- [13] Y. Liu, A. Tsyvinski, X. Wu, Common Risk Factors in Cryptocurrency, *J. Finance* 77 (2022), 1133–1177. <https://doi.org/10.1111/jofi.13119>.
- [14] S.M. Lundberg, S.-I. Lee, A Unified Approach to Interpreting Model Predictions, in: *Advances in Neural Information Processing Systems* 30, Curran Associates, Inc., pp. 4765–4774, (2017). <https://proceedings.neurips.cc/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html>.
- [15] A.J. McNeil, R. Frey, Estimation of Tail-Related Risk Measures for Heteroscedastic Financial Time Series: An Extreme Value Approach, *J. Empir. Finance* 7 (2000), 271–300. [https://doi.org/10.1016/S0927-5398\(00\)00012-8](https://doi.org/10.1016/S0927-5398(00)00012-8).

-
- [16] A.J. McNeil, R. Frey, P. Embrechts, Quantitative Risk Management: Concepts, Techniques and Tools, Revised ed., Princeton University Press, 2015.
- [17] T.J. Moskowitz, Y.H. Ooi, L.H. Pedersen, Time Series Momentum, *J. Financ. Econ.* 104 (2012), 228–250. <https://doi.org/10.1016/j.jfineco.2011.11.003>.
- [18] D.B. Nelson, Conditional Heteroskedasticity in Asset Returns: A New Approach, *Econometrica* 59 (1991), 347–370. <https://doi.org/10.2307/2938260>.
- [19] A. Neuhierl, R.T. Varneskov, Frequency Dependent Risk, *J. Financ. Econ.* 140 (2021), 644–675. <https://doi.org/10.1016/j.jfineco.2021.01.007>.
- [20] W.K. Newey, K.D. West, A Simple, Positive Semi-Definite, Heteroskedasticity and Autocorrelation Consistent Covariance Matrix, *Econometrica* 55 (1987), 703–708. <https://doi.org/10.2307/1913610>.
- [21] I. Welch, A. Goyal, A Comprehensive Look at the Empirical Performance of Equity Premium Prediction, *Rev. Financ. Stud.* 21 (2008), 1455–1508. <https://doi.org/10.1093/rfs/hhm014>.