DOI: https://doi.org/10.15276/hait.09.2026.35
UDC 004.8:004.89

Deep learning for cryptocurrency price forecasting: A taxonomy, benchmark, and gap analysis

Tien Duy Nguyen[1)]
ORCID: https://orcid.org/0009-0004-0732-2093; duynguyen.das@gmail.com.
The Vinh Tran[1)]
ORCID: https://orcid.org/0000-0002-4241-1065; vinhtt@ut.edu.vn. Scopus Author ID: 56835334700
Thanh Long Nguyen[1)]
ORCID: https://orcid.org/0009-0003-2438-1231; thanhlongnguyen.dev@gmail.com
1) Ho Chi Minh City University of Transport, 2, Vo Oanh Str. Ho Chi Minh City, 700000, Vietnam
*Corresponding author. E-mail: vinhtt@ut.edu.vn

ABSTRACT

Relevance. Cryptocurrency returns are volatile, non-stationary, and heavy-tailed. Purpose. This study develops a unified taxonomy and benchmark with reproducible scoring. **Tasks.** Representative architectures are compared and measured limitations are separated from proposed remedies. **Methods.** Ten statistical, linear, recurrent, attention-based, selective state-space, and pretrained models are evaluated on daily Bitcoin/Tether and Ethereum/Tether data from two thousand eighteen to two thousand twenty-six. Three expanding-window folds use horizons of five, ten, and twenty days. A target-history-only control shares post-warm-up calendar rows and normalization. Evaluation uses absolute error, symmetric percentage error, normalized-return directional agreement, and paired tests corrected for temporal dependence and multiple comparisons. **Scientific novelty**. The benchmark combines architectural coverage, aligned information-set control, and a zero forecast reference. **Practical significance**. It informs accuracy and resource trade-offs under weak predictive signal. **Results**. Best absolute-error skill over zero is approximately zero. Bitcoin autoregressive and recurrent model differences are not supported after correction. The recurrent model leads learned-model Ethereum mean absolute error, with many, but not all, pairwise advantages supported. Directional agreement remains near zero point five. TimesFM is the strongest native-input foundation reference; larger Chronos scale gives no consistent gain. Lag-Llama produces over-dispersed forecasts, and its low symmetric percentage error does not establish superior accuracy. **Conclusions.** This low-signal benchmark supports protocol-specific comparisons rather than exploitable return forecasts. Coverage is limited to two assets, three folds, and zero-shot foundation evaluation.

Keywords: cryptocurrency forecasting; deep learning; time-series forecasting; Transformers; Mamba; foundation models

For citation: **Nguyen T. D., Tran T. V.*, Nguyen T. L. "Deep learning for cryptocurrency price forecasting: A taxonomy, benchmark, and gap analysis". *Herald of Advanced Information Technology.* 2026; Vol.9 No.4: 567–587. DOI: https://doi.org/10.15276/hait.09.2026.35**

INTRODUCTION

The cryptocurrency market has evolved into one of the most active financial ecosystems, operating continuously through decentralized exchanges with substantial trading volume and liquidity. As cryptocurrencies become increasingly integrated into global financial systems, accurate price forecasting has become essential for algorithmic trading, portfolio management, risk assessment, and financial decision-making. However, cryptocurrency forecasting remains considerably more challenging than conventional financial forecasting because market prices exhibit non-stationarity, extreme volatility, heavy-tailed return distributions, nonlinear temporal dependencies, and frequent regime transitions. These statistical characteristics violate the assumptions of many conventional forecasting methods and require architectures capable of modeling complex temporal dynamics under rapidly changing market conditions [1], [2], [3], [4], [5].

To address these challenges, forecasting methodologies have evolved from statistical approaches such as ARIMA and GARCH to recurrent neural networks, Transformer-based architectures, selective State Space Models (SSMs), and, more recently, large foundation models for time-series forecasting. Each architectural generation introduces new sequence modeling mechanisms, including recurrent memory, self-attention, selective state updates, and large-scale pretraining, aiming to improve long-range dependency modeling and computational efficiency. Nevertheless, it remains unclear whether increasing architectural complexity consistently translates into superior forecasting performance under the highly volatile and non-stationary characteristics of cryptocurrency markets [6], [7], [8], [9].

Despite rapid methodological progress, several important limitations remain. Existing studies typically evaluate individual forecasting models or isolated architectural families using heterogeneous datasets, forecasting horizons, validation protocols, and evaluation metrics, making objective

© Nguyen D., Tran T. *, Nguyen L., 2026

This is an open access article under the CC BY license (https://creativecommons.org/licenses/by/4.0/deed.uk)

comparisons difficult. Furthermore, relatively few studies systematically investigate how cryptocurrency-specific characteristics – including heavy-tailed distributions,regime changes, and cross-asset dependencies – influence the strengths and limitations of different forecasting paradigms. Consequently, the relative effectiveness of modern forecasting architectures for cryptocurrency prediction remains insufficiently understood [3], [9], [10].

To address these limitations, this paper presents a taxonomy-driven review and a unified empirical benchmark of representative forecasting architectures for cryptocurrency prediction. Existing methods are organized by their sequence-modeling mechanisms, as detailed in the literature review. Representative architectures are evaluated under a common expanding-window walk-forward protocol using BTC/USDT and ETH/USDT data, with a separate target-history-only control to audit input-set sensitivity. The study combines literature review and empirical benchmarking to identify measured limitations and testable future directions.

The main contributions of this study are summarized as follows:

1) A unified architectural taxonomy that organizes cryptocurrency forecasting methods into five representative forecasting paradigms plus a lightweight linear baseline according to their sequence modeling mechanisms.

2) A standardized benchmark that evaluates representative statistical, deep learning, State Space, and foundation models under identical datasets, walk-forward validation, and evaluation metrics.

3) An evidence-driven gap analysis identifies heavy-tailed modeling, multi-task forecasting, cross-asset dependency learning, foundation model adaptation, and regime-aware evaluation as priorities for further testing.

Unlike surveys that compare published scores across heterogeneous datasets, this benchmark compares representative model families on the same assets, horizons, temporal partitions, and metrics. Its additional univariate control distinguishes information-set sensitivity from practical native-input performance. The contribution is a transparent cross-paradigm reference, rather than an exhaustive search or a claim of universally superior architecture.

LITERATURE REVIEW AND PROBLEM STATEMENT

Architectural Taxonomy of Cryptocurrency Forecasting. Recent advances in time-series forecasting have produced a diverse collection of architectures with substantially different modeling assumptions, computational characteristics, and sequence representation mechanisms. Existing surveys typically review individual forecasting models or focus on a single architectural family, making it difficult to understand the evolutionary relationship among modern forecasting paradigms and their suitability for cryptocurrency prediction [9], [10].

To provide a unified perspective, this study organizes cryptocurrency forecasting methods into five architectural paradigms plus a lightweight linear baseline, as illustrated in Fig. 1:

1) statistical and classical machine learning models;

2) recurrent neural networks;

3) transformer-based architectures;

4) selective state space models ;

5) foundation models.

Rather than following publication chronology, this taxonomy categorizes forecasting methods according to their underlying sequence modeling mechanisms. Each paradigm represents a distinct stage in the evolution of time-series forecasting, introducing different trade-offs between interpretability, computational efficiency, long-range dependency modeling, and transferability. The proposed taxonomy serves as the conceptual framework for both the literature review and the empirical benchmark presented later in this paper.

Problem Formulation. Cryptocurrency forecasting may target future closing prices C_t, log-returns $r_t = ln(C_t/C_{t-1})$, or the sign of price movement. These are distinct objectives. Although the title describes the application broadly as price forecasting, the empirical benchmark evaluates multi-horizon normalized log-returns.

For a 60-day input window, the model predicts up to 20 future normalized returns $z_t = (r_t - \mu train)/\sigma train$. Fold-specific means and standard deviations are fitted on training data only. Statistical and zero-shot foundation models forecast raw close-price paths; successive predicted prices are converted to log-returns and then to the same normalized target space.

Price-level error, return error, and directional agreement must be distinguished. Mean absolute error (MAE) is a dimensionless normalized-return error. Directional accuracy (DA) compares signs of normalized returns; centering shifts the zero threshold, so this is not strictly equivalent to raw price direction. Raw-return direction, class-aware baselines, and economic utility require separate evaluation.

ISSN 2663-0176 (Print)
ISSN 2663-7731 (Online)

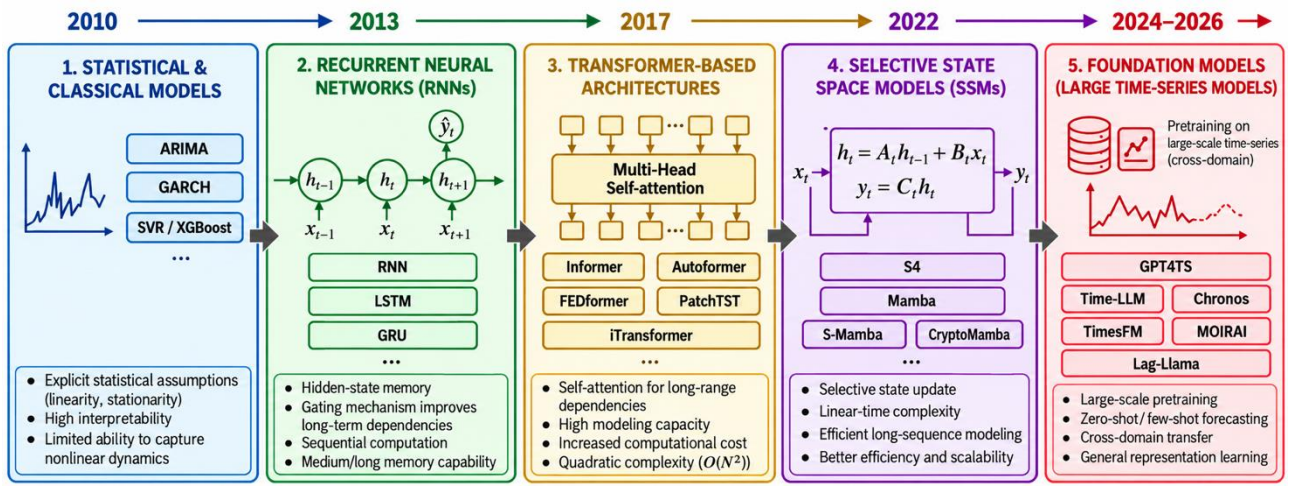


Fig. 1. **Evolution of forecasting architectures for cryptocurrency time-series prediction**

***Source*: compiled by the authors**

The benchmark standardizes assets, temporal partitions, horizons, and metrics while retaining each model family's native inputs. A separate univariate control gives all models close-derived target-history input. This improves information-set comparability but does not equate pretraining, training budgets, or target representations.

Evolution of Forecasting Architectures Fig. 1 illustrates the evolution of cryptocurrency forecasting architectures from statistical models to modern foundation models. The proposed taxonomy organizes existing approaches into five architectural paradigms plus a lightweight linear baseline based on their underlying sequence modeling principles, providing the foundation for the comparative benchmark presented in the Materials and Methods section. Statistical and Classical Machine Learning.

Statistical forecasting methods constitute the foundation of cryptocurrency forecasting. Representative approaches include ARIMA, GARCH, Support Vector Regression (SVR), Random Forests, and gradient boosting models [1], [2], [5], [11], [12]. Recent studies continue to report competitive performance of statistical and hybrid approaches for stock market forecasting [13]. These methods model temporal dynamics through predefined stochastic assumptions or handcrafted features, providing interpretable forecasts with low computational cost.

Despite their efficiency, statistical models generally assume relatively stable temporal structures and therefore struggle with non-stationarity, heavy-tailed distributions, nonlinear dependencies, and abrupt regime transitions that characterize cryptocurrency markets. Classical machine learning methods alleviate some of these limitations through nonlinear regression but remain heavily dependent on feature engineering and often exhibit limited robustness under changing market conditions [3], [4].

Recurrent neural networks, particularly LSTM and GRU, use gated hidden states to learn nonlinear temporal dependencies and are widely applied to time-series forecasting [14]. Their gated memory mechanisms substantially improved long-term dependency modeling and established LSTM as one of the strongest deep learning baselines for financial forecasting, including hybrid CNN–LSTM approaches for cryptocurrency price prediction [15].

However, sequential computation prevents efficient parallelization and limits scalability for long input sequences. Furthermore, recurrent memory gradually degrades over very long horizons, motivating the development of attention-based forecasting architectures capable of modeling global temporal interactions.

Transformer-based Architectures. Transformer architectures replace recurrent computation with self-attention, allowing direct interactions among distant observations while enabling efficient parallel training [16]. Representative forecasting models include Informer [17], Autoformer [18], FEDformer [19], Temporal Fusion Transformers [20], PatchTST [8], Crossformer [21], and iTransformer [22].

Compared with recurrent architectures, Transformers substantially improve long-range dependency modeling and have become the dominant paradigm for multivariate time-series forecasting. Nevertheless, quadratic attention complexity and sensitivity to distribution shifts remain important limitations, particularly for highly volatile cryptocurrency markets where temporal patterns evolve rapidly. Cryptocurrency-specific studies have further combined Transformer and

recurrent components with technical indicators to improve price prediction accuracy [23], [24].

Hybrid forecasting architectures combine complementary mechanisms. CNN–LSTM models use convolutional layers to extract local patterns before gated temporal aggregation [15]; attention-based Transformer–GRU models combine global interaction with recurrent memory [24]. Attention and state-space combinations likewise motivate efficient hybrid sequence modeling [25], [26]. These approaches are covered in the taxonomy but are not separately benchmarked here. Published results differ in assets, feature sets, horizons, and validation protocols, so they cannot establish a common ranking.

Selective State Space Models. SSMs provide an alternative sequence modeling paradigm by replacing quadratic attention with linear-time state-space representations [26], [7]. Mamba further extends this framework through input-dependent selective state updates, enabling adaptive memory while maintaining linear computational complexity.

These characteristics make SSMs particularly attractive for cryptocurrency forecasting, where long historical sequences, regime changes, and heavy-tailed behavior require efficient yet flexible temporal representations. Nevertheless, empirical evidence remains limited, especially regarding cross-asset dependency learning and standardized benchmark evaluation [25], [27], [28], [29], [30], [31].

Foundation Models. Foundation models extend forecasting through large-scale pretraining on heterogeneous temporal datasets [6], [32], [33], [34], [35], [36]. Chronos scales and quantizes numerical values into a fixed vocabulary and uses a T5-family encoder-decoder to generate probabilistic forecast samples [6]. TimesFM uses a decoder-only architecture for point forecasting [32]. Lag-Llama uses lagged covariates and a probabilistic decoder-only model [38]. MOIRAI uses a universal masked encoder supporting varying numbers of variables and forecasting horizons [35]. Scaling and quantization let Chronos reuse language-model token prediction for numerical sequences; its sampled continuations provide a predictive distribution whose median supplies the point forecast. This representation can lose numerical resolution and does not guarantee calibration under cryptocurrency distribution shifts.

GPT4TS adapts a pretrained language-model backbone with lightweight task-specific components [36], whereas Time-LLM reprograms time-series patches into the embedding space of a frozen language model [34]. These approaches differ from native numerical time-series pretraining. Here, Chronos, TimesFM, and Lag-Llama are evaluated only in zero-shot mode to measure off-the-shelf transfer without cryptocurrency parameter updates. Fine-tuning or domain adaptation would answer a separate question and requires controlled training budgets and temporal validation. LLMTime [33] represents numerical series as digit strings and prompts general-purpose language models for continuation without time-series-specific training.

Comparative Analysis and Research Gaps. The evolution of forecasting architectures demonstrates a continuous transition from model-driven statistical approaches toward increasingly data-driven representation learning. Rather than replacing previous paradigms, each architectural generation addresses specific limitations while introducing new computational trade-offs (Table 1).

Foundation models are grouped by their pretraining and transfer setting; their underlying architectures may overlap with Transformer or State Space paradigms.

Heterogeneous datasets and validation protocols make objective comparison difficult [9], [10], motivating the unified benchmark. Beyond evaluation, five challenges remain:

- G1. Heavy-tailed modeling. Conventional regression objectives may inadequately represent the pronounced tails of cryptocurrency returns [4].
- G2. Multi-task forecasting. Return, direction, volatility, and uncertainty are often modeled separately rather than jointly.
- G3. Cross-asset dependency learning. Independent asset models omit dynamic relationships among digital assets.
- G4. Foundation model adaptation. Zero-shot transfer under cryptocurrency-specific distributions requires controlled domain-adaptation tests [6], [34].
- G5. Regime-aware evaluation. Aggregate metrics leave robustness across bull, bear, sideways, and high-volatility periods unresolved.

These research gaps motivate the benchmark and subsequent analysis. The experiments directly measure forecasting errors, directional agreement, and temporal variation; proposed improvements from heavy-tail objectives, cross-asset learning, and domain adaptation remain hypotheses requiring dedicated tests.

ISSN 2663-0176 (Print)
ISSN 2663-7731 (Online)

Table 1. **Comparison of representative forecasting paradigms**

Paradigm	Long-range dependency	Complexity	Transferability	Typical representative
Statistical	Low	O(N)	Low	ARIMA
RNN	Medium	O(N)	Low	LSTM
Transformer	High	$O(N^2)$	Medium	PatchTST
SSM	High	O(N)	Medium	Mamba
Foundation	High	High	High	Chronos, TimesFM, Lag-Llama

***Source*: compiled by the authors**

RESEARCH AIM AND OBJECTIVES

Research Aim. The primary aim of this study is to provide a unified understanding of modern cryptocurrency forecasting architectures through a taxonomy-driven literature review and a standardized empirical benchmark. By evaluating representative forecasting paradigms under a common experimental framework, this study seeks to identify their relative strengths, limitations, and unresolved research challenges, thereby providing practical guidance for the development of next-generation cryptocurrency forecasting systems.

Research Objectives. To achieve this aim, the study pursues the following objectives:

- O1. Taxonomy Development: Develop a unified taxonomy that organizes cryptocurrency forecasting methods into five representative architectural paradigms plus a lightweight linear baseline.
- O2. Benchmark Evaluation: Compare ten representative models using common assets, horizons, temporal partitions, and metrics, and audit information-set sensitivity with a separate target-history-only benchmark.
- O3. Architectural Analysis: Analyze the strengths and limitations of different forecasting paradigms under cryptocurrency-specific market characteristics.
- O4. Research Gap Identification: Identify unresolved challenges and future research directions for next-generation cryptocurrency forecasting architectures

Research Questions. Based on these objectives, the study addresses the following research questions:

- RQ1. How do representative forecasting paradigms compare under a unified benchmarking framework?
- RQ2. Does increasing architectural complexity consistently improve forecasting performance in cryptocurrency markets?
- RQ3. Do selective State Space Models provide greater robustness than Transformer-based architectures under highly volatile market conditions?
- RQ4. To what extent can foundation models generalize to cryptocurrency forecasting without domain-specific adaptation?
- RQ5. What research gaps and future directions should be prioritized for next-generation cryptocurrency forecasting systems?

MATERIALS AND METHODS

Dataset and Market Characteristics. This study employs daily BTC/USDT and ETH/USDT data collected from Binance via ccxt. The single pinned snapshot used for both descriptive analysis and benchmarking spans 1 January 2018 to 24 May 2026 and contains 3,066 daily Open, High, Low, Close, and Volume (OHLCV) rows per asset. Bitcoin and Ethereum were selected for their long data histories and distinct market dynamics. This two-asset design does not represent smaller, less liquid cryptocurrencies or the entire market.

The raw variables are the five OHLCV fields. Close prices are transformed into the normalized log-return target described above. Each asset has 3,065 close-derived log-return observations. Removing the first 33 rows of the raw daily series during feature preparation leaves 3,033 usable feature rows, yielding 2,954 full-series sliding windows with a 60-day lookback and a 20-day output horizon. The full-series window count is descriptive; benchmark windows are constructed separately within each training, validation, and test partition. Raw bars, feature rows, and overlapping samples are distinct counts; Table 2 reports raw daily observations.

Both BTC and ETH exhibit pronounced non-stationarity in price levels, while first-order differencing produces stationary return series. In addition, their high kurtosis values indicate heavy-tailed return distributions, confirming that cryptocurrency markets deviate substantially from Gaussian assumptions. These characteristics highlight the need for forecasting models capable of

handling distributional shifts, extreme market movements, and evolving market conditions.

Fig. 2 shows the market phases covered by the unified snapshot and the three expanding-window folds (0-2), whose dates, 60-day gaps, and window counts are detailed in Table 6. Although these folds preserve temporal order, their limited number and the absence of regime-specific evaluation prevent firm conclusions about robustness to regime changes. Fig. 2 shows the actual expanding-window folds; exact dates, 60-day gaps, and forecast-window counts are specified in Table 6. Folds 0-2 follow the chronological runner indices.

Table 3 and Fig. 3 characterize the heavy tails of daily log-returns in the same pinned snapshot through 24 May 2026. Pearson kurtosis reaches 20.8156 for BTC and 15.7489 for ETH, while absolute returns above 10 % occur more frequently than under the fitted Gaussian references shown in Fig. 3, motivating explicit tail-risk evaluation.

Table 4 and Fig. 4 further show that cross-asset dependencies are dynamic and regime-dependent. The 60-day rolling Pearson correlation between BTC and ETH averages 0.84, falling to 0.78 during the April 2020-November 2021 bull market, rising to 0.89 during the December 2021-December 2022 bear market and crash, and moderating to 0.81 during the November 2023-May 2026 bull market expansion. These findings indicate that inter-asset relationships are time-varying rather than stationary; suggesting that independently trained models may overlook valuable shared market information. This observation motivates future research on cross-asset representation learning and multi-task forecasting.

Table 2. **Summary statistics of the cryptocurrency benchmark dataset (2018-2026)**

Statistic	BTC/USDT	ETH/USDT	Interpretation
Observation period	1 Jan 2018 – 24 May 2026	1 Jan 2018 – 24 May 2026	Unified pinned daily snapshot
Raw daily observations	3,066	3,066	Daily OHLCV bars
Mean daily log-return (%)	0.0570	0.0336	Average daily return
Standard deviation (%)	3.4514	4.5402	Daily return volatility
Minimum log-return (%)	−50.2607	−59.0534	Extreme downside
Maximum log-return (%)	17.8449	23.3750	Extreme upside
Skewness	−1.1384	−0.9389	Left skew
Excess kurtosis	17.8156	12.7489	Gaussian baseline = 0
ADF p-value (Price)	0.717570	0.366019	Unit root not rejected
ADF p-value (Log Return)	<0.001	<0.001	Stationary after log differencing

***Source*: compiled by the authors**

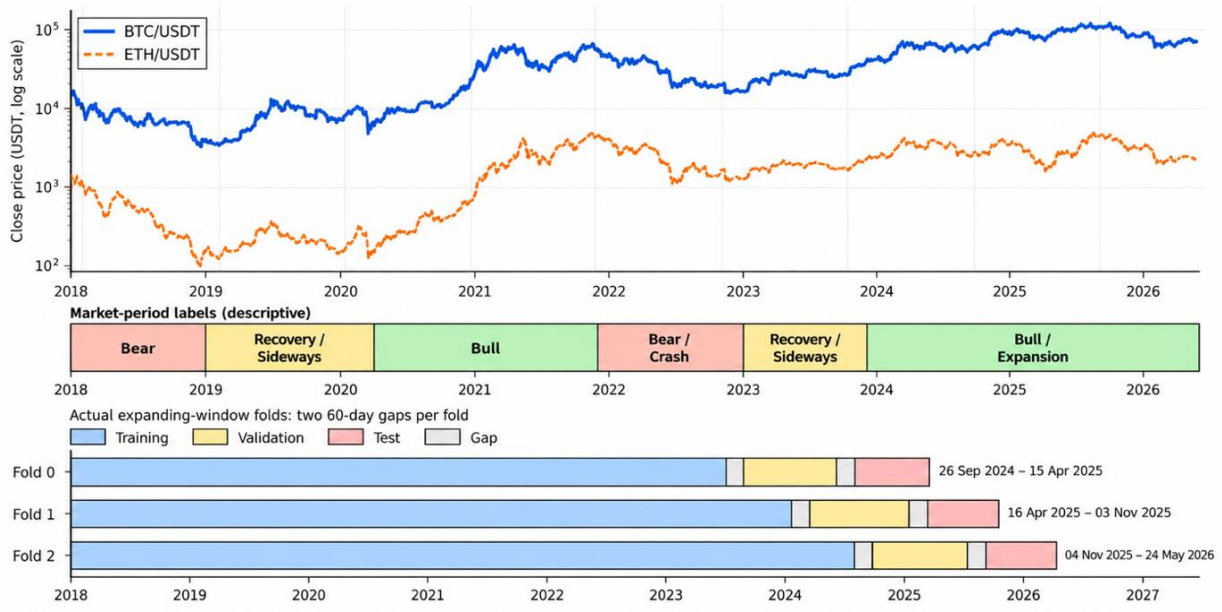


Fig. 2. **Cryptocurrency market evolution and actual benchmark folds (1 January 2018-24 May 2026)**
***Source*: compiled by the authors**

Table 3. **Statistical characteristics of BTC/USDT and ETH/USDT daily log-return distributions**

Statistic	BTC daily log-return	Interpretation	ETH daily log-return	Interpretation
Pearson kurtosis	20.8156	Heavy tails	15.7489	Heavy tails
Probability (|r| > 10 %)	1.99 %	Gaussian ≈ 0.38 %	4.11 %	Gaussian ≈ 2.76 %

***Source*: compiled by the authors**

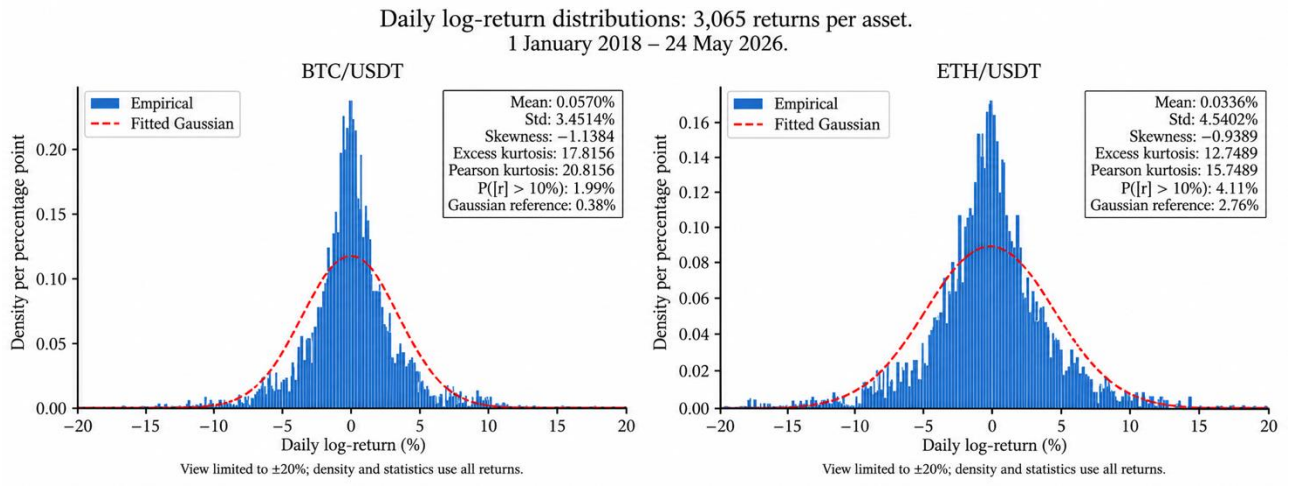


Fig.3. **Empirical distributions of BTC/USDT and ETH/USDT daily log-returns with fitted Gaussian curves (snapshot through 24 May 2026)**

***Source*: compiled by the authors**

Table 4. **Rolling correlation statistics between BTC/USDT and ETH/USDT daily log-returns**

Period (Start-End)	Market Regime	Mean Corr.	Std Dev.	Min Corr.	Max Corr.
Jan 2018 – Dec 2018	Bear Market	0.88	0.05	0.75	0.94
Jan 2019 – Mar 2020	Recovery / Sideways	0.85	0.06	0.68	0.96
Apr 2020 – Nov 2021	Bull Market	0.78	0.09	0.53	0.97
Dec 2021 – Dec 2022	Bear Market / Crash	0.89	0.05	0.77	0.97
Jan 2023 – Oct 2023	Recovery / Sideways	0.88	0.05	0.74	0.97
Nov 2023 – 24 May 2026	Bull Market / Expansion	0.81	0.08	0.64	0.95
Overall (Jan 2018 – 24 May 2026)		0.84	0.08	0.53	0.97

***Source*: compiled by the authors**

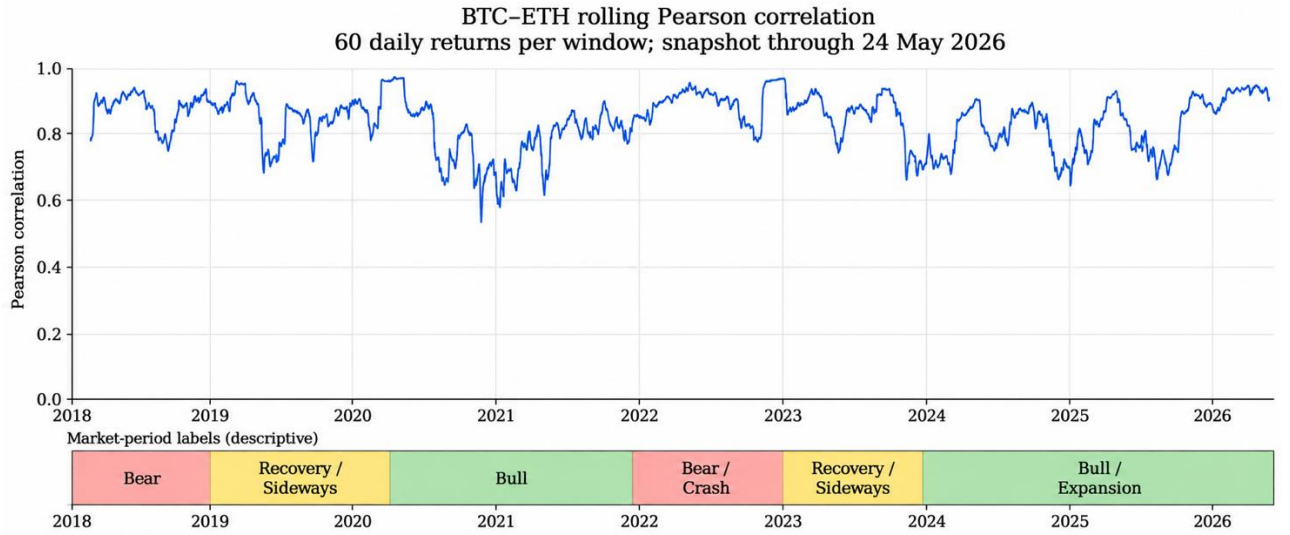


Fig. 4. **BTC-ETH rolling correlation (snapshot through 24 May 2026)**

***Source*: compiled by the authors**

Feature Engineering. The five OHLCV channels are augmented by 18 engineered variables, yielding 23 channels: SMA-20, EMA-12, EMA-26, MACD, MACD signal, MACD histogram, RSI-14, ROC-10, Stochastic %K and %D, Bollinger upper and lower bands, Bollinger width and %B, ATR-14, OBV, CMF-20, and close log-return. Trend, momentum, volatility, and volume features provide complementary market-state information. Indicators follow the ta library definitions and are computed causally; missing warm-up rows are removed before splitting. Non-target channels are standardized using training-fold statistics only.

Price-level channels, including OHLC, moving averages, and Bollinger bands, remain non-stationary after fold-based standardization. Test-period levels may exceed the training distribution. This is a plausible source of input sensitivity, not a causal explanation established by the aligned control; relative-price or return-based features require a separate test.

For closing price C, high H, low L, and volume V, SMA_n is the mean of the last n closes and EMA_n uses $\alpha = 2/(n+1)$. $MACD = EMA_{12} - EMA_{26}$; its signal is $EMA_9(MACD)$, and its histogram is MACD minus signal. $RSI_{14} = 100 - 100/(1 + RS)$, where RS is the ratio of exponentially smoothed positive and negative close changes. $ROC_{10} = 100(C_t/C_{t-10} - 1)$. Stochastic $\%K = 100(C_t - min_{14} L)/(max_{14} H - min_{14} L)$, and $\%D$ is its three-period mean.

Bollinger bands are $U = SMA_{20} + 2s_{20}$ and $B = SMA_{20} - 2s_{20}$; $width = 100(U - B)/SMA_{20}$ and $\%B = (C - B)/(U - B)$. True range is $max(H - L, |H - Cprevious|, |L - Cprevious|)$; ATR_{14} is its smoothed average. OBV accumulates volume positively for rising closes and negatively for falling closes. $CMF_{20} = \Sigma_{20}[V(2C - H - L)/(H - L)]/\Sigma_{20}V$, with zero-range handling following the library implementation. Log-return is $ln(C_t/C_{t-1})$.

Benchmark Models. The benchmark covers the principal forecasting paradigms reviewed above: ARIMA(5,1,0) [11], LSTM [14], PatchTST [8], iTransformer [22], vanilla Mamba [7], Chronos-T5-Small and Chronos-T5-Large [6], TimesFM [32], and Lag-Llama [38]. DLinear [37] provides a lightweight baseline for assessing whether compact models remain competitive with larger architectures. Model selection prioritizes architectural coverage over numerous variants within each family. Table 5 summarizes their parameters, inputs, forecasting strategies, and roles.

ARIMA is refitted by maximum likelihood on each test window's 60-day raw-close history. The five supervised neural models are trained from scratch within each fold, producing 20-step forecasts evaluated at $H \in \{5, 10, 20\}$. LSTM uses a three-layer encoder-decoder with hidden dimension 128, dropout 0.1, and autoregressive decoding. PatchTST uses patch length 16, stride 8, three encoder layers, width 128, and eight attention heads. iTransformer represents each of the 23 variates as a token. Mamba uses two selective-state blocks, width 128, state dimension 16, convolution width 4, and expansion factor 2. DLinear uses a moving-average kernel of 25 and two shared 60-to-20 projections, totaling 2,440 parameters. These projections operate independently across channels; selecting the close-channel output makes its effective input solely the normalized close-return history.

The four foundation models use 60-day raw-close histories without cryptocurrency fine-tuning. Chronos uses temperature 1.0, top-k 50, top-p 1.0, 20 sample paths, and batches of up to 64 windows. Its PyTorch median selects the lower central sample. Lag-Llama uses Student-t sampling, 100 paths, batch size 1, and nonnegative price clamping before feedback; its NumPy median averages the two central samples. Both convert median price paths, anchored at the last observed close, into successive log-returns using a price floor of 10^{-8}, then normalize them with the training-fold scaler. TimesFM produces point forecasts. Chronos-Small, Chronos-Large, and TimesFM contain 46, 710, and 200 million parameters, respectively. Seed 42 applies only to supervised initialization and shuffling. Chronos/Lag-Llama sampling was unseeded, and sample paths and random-generator states were not retained. Saved forecasts and targets permit metric recomputation, but repeated-seed sampling variability remains unquantified.

In the native protocol, supervised models receive 23-channel tensors, although DLinear's selected output depends only on close returns; ARIMA and foundation models receive raw-close histories. The aligned control selects one close-only channel after the same indicator warm-up filter, preserving calendar rows, folds, lookback, horizons, and normalization. Its 60 runs—10 models × two assets × three folds—each save prediction and target arrays of shape (123, 20). This control assesses additional-channel sensitivity where effective inputs change, while retaining training and pretraining differences. Chronos/Lag-Llama results are separate stochastic realizations with unchanged input formulations and cannot establish information-set effects. TimesFM uses identical inputs and configurations across protocols and reports identical metrics.

ISSN 2663-0176 (Print)
ISSN 2663-7731 (Online)

Table 5. **Representative forecasting models included in the benchmark**

Model	Paradigm	Parameters	Input	Forecasting strategy	Learning setting	Role in benchmark
ARIMA(5,1,0)	Statistical	~5	Close-derived target	Price path → normalized returns	Refitted per test window	Classical baseline
DLinear	Lightweight DL	2,440	Normalized close-return history	Direct multi-output	Supervised	Low-complexity baseline
LSTM	Recurrent	409,729	OHLCV + indicators	Autoregressive encoder-decoder	Supervised	Sequential modeling
PatchTST	Transformer	613,396	OHLCV + indicators	Direct multi-output	Supervised	Patch-based attention
iTransformer	Transformer	605,460	OHLCV + indicators	Direct multi-output	Supervised	Variable-wise attention
Mamba	Selective SSM	356,116	OHLCV + indicators	Direct multi-output	Supervised	Selective memory modeling
Chronos-T5-Small	Foundation	46 M	Close-derived target	Native multi-step	Zero-shot	Foundation baseline
Chronos-T5-Large	Foundation	710 M	Close-derived target	Native multi-step	Zero-shot	Scale-effect verification
TimesFM	Foundation	200 M	Close history	Point price path → normalized returns	Zero-shot	Decoder-only reference
Lag-Llama	Foundation	≈2.45M	Close history	Sample-median price path → normalized returns	Zero-shot	Probabilistic reference

Source: **compiled by the authors**

Experimental Design

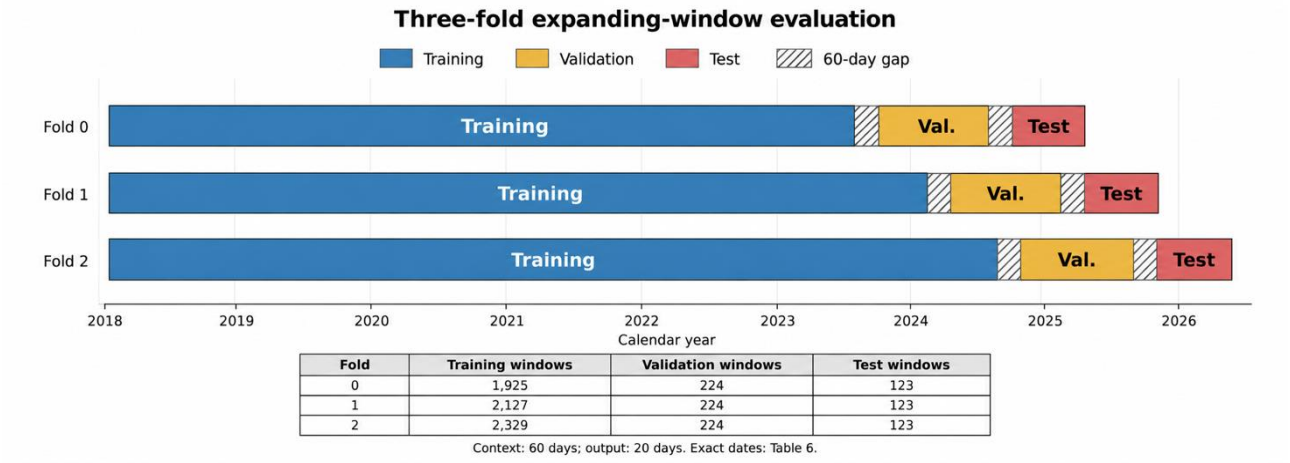
Walk-forward Evaluation Protocol. The study uses three expanding-window folds to preserve temporal order (Fig. 5; Table 6). Supervised neural models are trained separately within each fold, ARIMA is refitted on each historical context, and foundation models are evaluated zero-shot. All models share the same calendar partitions and forecasting horizons of 5, 10, and 20 days. Two 60-day gaps separate training, validation, and testing.

Scaling parameters are estimated exclusively from training rows. Sliding windows are constructed entirely within each partition using a 60-day context and a 20-day output, yielding (N-79) windows from (N) rows. Shorter horizons are evaluated by slicing the same 20-step forecasts, ensuring identical forecast origins across horizons. Each test partition provides 123 windows; its first 60 days supply the initial context. Serial dependence from overlapping windows is accounted for in the paired-test procedure described below.

Table 6. **Calendar boundaries, row counts, and usable windows for the three expanding folds**

Fold	Partition	Start date	End date	Rows	Windows
0	Training	2018-02-03	2023-07-30	2,004	1,925
0	Validation	2023-09-29	2024-07-27	303	224
0	Test	2024-09-26	2025-04-15	202	123
1	Training	2018-02-03	2024-02-17	2,206	2,127
1	Validation	2024-04-18	2025-02-14	303	224
1	Test	2025-04-16	2025-11-03	202	123
2	Training	2018-02-03	2024-09-06	2,408	2,329
2	Validation	2024-11-06	2025-09-04	303	224
2	Test	2025-11-04	2026-05-24	202	123

Source: **compiled by the authors**



Fold	Training windows	Validation windows	Test windows
0	1,925	224	123
1	2,127	224	123
2	2,329	224	123

Context: 60 days; output: 20 days. Exact dates: Table 6.

Fig. 5. **Three-fold expanding-window walk-forward validation**
***Source*: compiled by the authors**

Training Configuration. Supervised neural models use Table 7 and an NVIDIA RTX 5070 Ti GPU with 16.6 GB VRAM. Each protocol comprises 60 canonical runs (ten models × two assets × three folds): 30 supervised neural runs, six ARIMA runs, and 24 zero-shot foundation runs. Horizons slice the same 20-step outputs. Training wall-clock time, epochs to early stopping, training peak memory, processor utilization, and energy consumption are not reported.

Reproducibility and Fairness. Forecasting performance is reported using MAE, symmetric mean absolute percentage error (SMAPE), and DA on normalized log-returns. MAE is the primary reported point-error criterion; SMAPE is a secondary diagnostic because centered returns and forecast dispersion can distort its denominator. Root mean squared error (RMSE) is computed by the evaluation pipeline but is not included in the compact result tables and is not used for table-based ranking claims.

For n forecasted observations, MAE is defined as

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \,. \qquad (1)$$

Here n is the number of evaluated forecast entries; y_i and $\hat{y}_i$ denote observed and predicted normalized returns.

SMAPE is defined as

$$SMAPE = \frac{1}{n}\sum_{i=1}^{n}\frac{2|y_i-\hat{y}_i|}{|y_i|+|\hat{y}_i|+2\varepsilon}. \qquad (2)$$

Here $\varepsilon = 10^{-8}$ prevents division by zero. The table reports SMAPE as a percentage, i.e., 100 times the ratio in formula (2). Near-zero normalized returns can still yield large SMAPE, so a lower value does not imply better calibration or trading utility.

For multi-horizon return forecasting, DA is computed as

$$DA = \frac{1}{n}\sum_{i=1}^{n}\mathbf{1}\,[sign(y_i) = sign(\hat{y}_i)]. \qquad (3)$$

The indicator equals one when the signs of observed and predicted normalized returns agree. DA around 0.5 indicates weak directional information under this definition; 0.5 is a heuristic baseline rather than a tested null adjusted for class imbalance. Centering shifts the threshold relative to raw-return direction, and DA does not measure profitability.

Final results are reported as mean and standard deviation across the three walk-forward folds:

$$\bar{m} = \frac{1}{K}\sum_{k=1}^{K} m_k \,;$$
$$s_m = \sqrt{\frac{\sum_{k=1}^{K}(m_k-\bar{m})^2}{K-1}}; \quad K = 3. \qquad (4)$$

Here m_k is the metric in fold k. Standard deviation describes temporal variation across folds, not variation across random seeds. Lower MAE indicates lower point error; higher DA indicates greater normalized-return sign agreement. SMAPE is interpreted diagnostically.

The zero forecast sets every normalized-return prediction to zero. Its MAE equals $E|z_test|$; MAE skill is $1 - MAE_model/MAE_zero$. Zero in the centered target space corresponds to the training-fold mean raw return, rather than exactly zero raw return. The baseline is not included among the ten model pairs in the 270-test families. Paired inference uses per-window MAE differences, a Diebold–Mariano statistic with Newey–West heteroskedasticity and autocorrelation consistent variance at $lag\ H - 1$, and the Harvey–Leybourne–Newbold finite-sample correction. Holm adjustment across 270 aggregate model-pair tests per

ISSN 2663-0176 (Print)
ISSN 2663-7731 (Online)

protocol is the primary criterion at 0.05. Holm-adjusted Wilcoxon signed-rank tests provide a robustness check; Benjamini–Hochberg adjustments are also computed.

Table 7. Shared training hyperparameters for supervised neural models

Hyper-parameter	Value	Rationale
Optimizer	AdamW	Weight decay regularization
Learning rate	1×10^{-4}	Conservative; stable across architectures
Weight decay	1×10^{-5}	Light regularization
Batch size	64	Used for supervised neural-model training
Max epochs	100	Upper bound; early stopping active
Early stopping patience	15 epochs	Validation MSE criterion
Gradient clip norm	1.0	Prevents exploding gradients
LR schedule	Cosine annealing	Smooth decay over max epochs
Random seed	42	Controls supervised model initialization and data shuffling

***Source*: compiled by the authors**

RESEARCH RESULTS

Benchmark Results. Tables 8 and 9 report native-input results for BTC/USDT and ETH/USDT; Tables 11 and 12 report the aligned univariate control on the same assets. Table 10 presents MAE skill relative to the zero-forecast baseline for the model with the lowest mean MAE in each asset–horizon–protocol setting. Both protocols share calendar rows, folds, horizons, and normalization. MAE and RMSE are dimensionless normalized-return errors; SMAPE is expressed as a percentage. DA measures sign agreement in normalized-return space and should not be interpreted directly as price-direction accuracy.

Statistical comparison. Holm-adjusted DM-HAC-HLN tests identify significant differences in 173 of 270 native-input comparisons and 171 of 270 aligned-univariate comparisons. Holm-adjusted Wilcoxon tests yield the same numbers of significant comparisons. No significant ARIMA-LSTM differences are detected on BTC at H = 5, 10, or 20. ARIMA, LSTM, and PatchTST form a competitive group for BTC MAE, without a consistently established winner. LSTM achieves the lowest mean MAE among learned models on ETH, although its pairwise advantages are not uniformly significant. These tests exclude the zero baselines and therefore do not establish superiority to it.

Table 8. **Native-input forecasting results for BTC/USDT over three walk-forward folds**
Mean ± standard deviation over three walk-forward folds

Model	H5 MAE	H10 MAE	H20 MAE	H5 SMAPE	H20 SMAPE	H5 DA	H20 DA
ARIMA(5,1,0)	0.458658 ± 0.098898	0.459071 ± 0.097016	0.463714 ± 0.098209	176.1±1.9	178.4±3.2	0.519±0.037	0.528±0.012
DLinear	0.478309 ± 0.104106	0.478426 ± 0.101108	0.483105 ± 0.103219	157.2±2.2	158.7±1.5	0.506±0.014	0.502±0.005
LSTM	0.457381 ± 0.097438	0.458681 ± 0.096641	0.464261 ± 0.098652	178.1±10.8	187.4±2.7	0.537±0.017	0.491±0.006
PatchTST	0.459936 ± 0.095933	0.464824 ± 0.096863	0.470127 ± 0.099032	160.1±2.3	163.2±2.9	0.540±0.011	0.511±0.007
iTransformer	0.472704 ± 0.093028	0.474520 ± 0.090644	0.479257 ± 0.094957	159.4±1.6	160.6±1.4	0.512±0.005	0.505±0.009
Mamba	0.496204 ± 0.083113	0.498880 ± 0.079986	0.500908 ± 0.078080	154.7±5.4	153.7±3.0	0.498±0.011	0.519±0.017
Chronos-Small	0.502865 ± 0.101530	0.508784 ± 0.104165	0.527152 ± 0.107405	157.1±2.8	154.5±2.0	0.489±0.021	0.503±0.026
Chronos-Large	0.508382 ± 0.116055	0.520663 ± 0.123140	0.539250 ± 0.126411	155.2±3.3	154.0±2.4	0.512±0.017	0.505±0.001
TimesFM	0.474958 ± 0.107372	0.469802 ± 0.108451	0.473412 ± 0.110379	160.4±3.4	164.1±4.0	0.501±0.041	0.520±0.037
Lag-Llama	1.116525 ± 0.144022	1.074149 ± 0.082177	1.060265 ± 0.058072	154.5±4.3	152.9±2.6	0.494±0.013	0.500±0.009
Zero forecast	0.458056 ± 0.097974	0.458974 ± 0.096845	0.464013 ± 0.098473	200.0±0.0	200.0±0.0	n/a	n/a

***Source*: compiled by the authors**

Table 9. **Native-input forecasting results for ETH/USDT over three walk-forward folds**

Model	H5 MAE	H10 MAE	H20 MAE	H5 SMAPE	H20 SMAPE	H5 DA	H20 DA
ARIMA(5,1,0)	0.595942 ± 0.007380	0.581224 ± 0.012659	0.577046 ± 0.025334	168.3±3.6	183.2±3.3	0.472±0.002	0.498±0.020
DLinear	0.590631 ± 0.003045	0.592160 ± 0.001513	0.590949 ± 0.015735	156.4±4.1	156.1±2.3	0.489±0.033	0.501±0.024
LSTM	0.567597 ± 0.013945	0.567761 ± 0.015848	0.571707 ± 0.026408	174.0±8.1	178.1±12.1	0.502±0.017	0.484±0.022
PatchTST	0.578481 ± 0.009968	0.578749 ± 0.008625	0.580175 ± 0.020872	165.7±1.6	164.5±1.3	0.496±0.006	0.491±0.018
iTransformer	0.576024 ± 0.012496	0.574783 ± 0.012000	0.578123 ± 0.023602	164.6±3.1	163.8±1.3	0.495±0.001	0.503±0.011
Mamba	0.576959 ± 0.006918	0.574070 ± 0.008087	0.577938 ± 0.020370	162.7±5.1	163.7±3.3	0.495±0.024	0.499±0.024
Chronos-Small	0.634471 ± 0.011074	0.635388 ± 0.008743	0.657619 ± 0.023149	158.8±3.4	153.8±3.2	0.496±0.016	0.508±0.027
Chronos-Large	0.631389 ± 0.013340	0.645732 ± 0.017244	0.667553 ± 0.019328	156.6±3.2	152.1±1.3	0.492±0.010	0.513±0.019
TimesFM	0.598707 ± 0.002823	0.592161 ± 0.010693	0.591054 ± 0.020666	163.6±7.5	166.6±3.2	0.491±0.023	0.496±0.012
Lag-Llama	1.061837 ± 0.132364	0.939119 ± 0.098844	0.877781 ± 0.094679	150.6±1.1	149.8±0.8	0.497±0.007	0.495±0.006
Zero forecast	0.566819 ± 0.013889	0.565748 ± 0.015857	0.569247 ± 0.027334	200.0±0.0	200.0±0.0	n/a	n/a

***Source*: compiled by the authors**

Table 10. **Best-model MAE skill relative to the zero forecast, by protocol**

Protocol	Asset	H5 skill	H10 skill	H20 skill
Native input	BTC/USDT	+0.15 %	+0.06 %	+0.06 %
Native input	ETH/USDT	−0.14 %	−0.36 %	−0.43 %
Aligned univariate	BTC/USDT	−0.05 %	−0.02 %	+0.06 %
Aligned univariate	ETH/USDT	−0.12 %	−0.09 %	−0.08 %

***Source*: compiled by the authors**

Table 11. **Aligned univariate control results for BTC/USDT (mean ± SD over three folds)**

Model	H5 MAE	H10 MAE	H20 MAE	H5 SMAPE	H20 SMAPE	H5 DA	H20 DA
ARIMA(5,1,0)	0.458658 ± 0.098898	0.459071 ± 0.097016	0.463714 ± 0.098209	176.1±1.9	178.4±3.2	0.519±0.037	0.528±0.012
DLinear	0.478309 ± 0.104106	0.478426 ± 0.101108	0.483105 ± 0.103219	157.2±2.2	158.7±1.5	0.506±0.014	0.502±0.005
LSTM	0.458271 ± 0.097748	0.459375 ± 0.096773	0.464444 ± 0.098489	189.2±0.4	189.9±1.5	0.496±0.033	0.491±0.038
PatchTST	0.459773 ± 0.095610	0.464987 ± 0.096722	0.470399 ± 0.098973	160.8±2.2	163.8±3.0	0.541±0.013	0.508±0.012
iTransformer	0.467339 ± 0.095514	0.467931 ± 0.092563	0.474471 ± 0.096361	159.9±1.5	160.8±2.1	0.507±0.013	0.509±0.014
Mamba	0.461767 ± 0.100764	0.462602 ± 0.096392	0.467302 ± 0.098069	171.3±0.3	172.9±1.7	0.489±0.010	0.495±0.010
Chronos-Small	0.492808 ± 0.101912	0.509090 ± 0.106652	0.530553 ± 0.118733	153.1±2.8	153.9±1.6	0.524±0.015	0.509±0.013
Chronos-Large	0.508964 ± 0.107431	0.509900 ± 0.114334	0.530612 ± 0.121518	156.2±1.5	152.8±2.0	0.500±0.014	0.515±0.006
TimesFM	0.474958 ± 0.107372	0.469802 ± 0.108451	0.473412 ± 0.110379	160.4±3.4	164.1±4.0	0.501±0.041	0.520±0.037
Lag-Llama	1.138432 ± 0.101598	1.080187 ± 0.089749	1.058367 ± 0.061517	154.4±2.6	153.0±1.0	0.489±0.017	0.499±0.010
Zero forecast	0.458056 ± 0.097974	0.458974 ± 0.096845	0.464013 ± 0.098473	200.0±0.0	200.0±0.0	n/a	n/a

***Source*: compiled by the authors**

ISSN 2663-0176 (Print)
ISSN 2663-7731 (Online)

Table 12. **Aligned univariate control results for ETH/USDT (mean ± SD over three folds)**

Model	H5 MAE	H10 MAE	H20 MAE	H5 SMAPE	H20 SMAPE	H5 DA	H20 DA
ARIMA(5,1,0)	0.595942 ± 0.007380	0.581224 ± 0.012659	0.577046 ± 0.025334	168.3±3.6	183.2±3.3	0.472±0.002	0.498±0.020
DLinear	0.590631 ± 0.003045	0.592160 ± 0.001513	0.590949 ± 0.015735	156.4±4.1	156.1±2.3	0.489±0.033	0.501±0.024
LSTM	0.567524 ± 0.013802	0.566232 ± 0.016000	0.569699 ± 0.027604	188.4±4.5	189.8±7.0	0.499±0.023	0.482±0.010
PatchTST	0.578378 ± 0.010409	0.579103 ± 0.009224	0.580290 ± 0.021315	166.4±1.8	164.6±1.3	0.494±0.000	0.489±0.015
iTransformer	0.580245 ± 0.007517	0.577911 ± 0.010450	0.581148 ± 0.019207	163.8±2.7	162.6±1.9	0.489±0.020	0.491±0.026
Mamba	0.571090 ± 0.012286	0.569350 ± 0.013425	0.574193 ± 0.025194	174.0±3.4	174.9±2.4	0.493±0.033	0.481±0.012
Chronos-Small	0.629598 ± 0.010079	0.635135 ± 0.013977	0.660424 ± 0.022278	157.1±2.8	154.0±1.7	0.494±0.005	0.500±0.012
Chronos-Large	0.626894 ± 0.015389	0.644209 ± 0.007586	0.666797 ± 0.010782	157.1±2.8	152.7±1.6	0.494±0.017	0.504±0.017
TimesFM	0.598707 ± 0.002823	0.592161 ± 0.010693	0.591054 ± 0.020666	163.6±7.5	166.6±3.2	0.491±0.023	0.496±0.012
Lag-Llama	1.065780 ± 0.128540	0.947151 ± 0.109479	0.880909 ± 0.093374	149.6±3.6	149.1±2.0	0.498±0.008	0.503±0.009
Zero forecast	0.566819 ± 0.013889	0.565748 ± 0.015857	0.569247 ± 0.027334	200.0±0.0	200.0±0.0	n/a	n/a

***Source*: compiled by the authors**

The results reveal five principal empirical observations:

- Observation 1. Overall forecasting performance. No model dominates both assets and all metrics. Across assets, horizons, and protocols, the best-model MAE skill relative to zero ranges from −0.43% to +0.15%. These negligible numerical gains suggest limited forecasting skill in the evaluated setting, despite differences in inter-model rankings. The reported pairwise tests exclude zero and therefore do not establish superiority to this baseline; nonsignificant model-pair differences also do not establish equivalence.
- Observation 2. Complexity versus accuracy. Compact and recurrent models remain competitive with larger architectures. DLinear provides a useful baseline with few parameters, while LSTM performs strongly by MAE, particularly on ETH. Increased model capacity offers no consistent accuracy advantage, suggesting that architectural complexity should be justified by measured improvements rather than parameter count alone.
- Observation 3. Selective State Space Model performance. Mamba shows asset- and metric-dependent performance without consistent MAE leadership. In the aligned control, it follows the leading BTC ARIMA/LSTM group and achieves the second-lowest learned-model ETH MAE, while remaining worse than zero. Its H5 MAE decreases from 0.496 to 0.462 on BTC and from 0.577 to 0.571 on ETH. These changes indicate protocol sensitivity but do not establish a causal benefit from removing indicators.
- Observation 4. Foundation-model transferability. TimesFM achieves the lowest native-input MAE among the evaluated foundation models, whereas increasing Chronos scale provides no consistent benefit. Lag-Llama combines relatively high MAE with low SMAPE. Because SMAPE depends on prediction and target magnitudes, especially near zero, this result alone does not establish superior point forecasting. These findings concern zero-shot inference and do not assess cryptocurrency-specific adaptation.
- Observation 5. Directional prediction capability. DA remains around 0.5, with maxima of 0.540 in the native protocol and 0.541 in the aligned control, both achieved by PatchTST on BTC at H5. DA measures sign agreement in centered normalized-return space. Without appropriate directional baselines and significance tests, these values do not establish reliable raw-price direction prediction or trading utility.

Information-set alignment preserves common evaluation targets, calendar windows, folds, and normalization, but does not equalize training procedures, internal forecast representations, or pretraining budgets. Chronos and Lag-Llama retain

the same raw-close input formulation across protocols; their score changes reflect separate unseeded sampling realizations and cannot identify input effects.

Comparative Analysis Across Forecasting Paradigms

Table 13 summarizes architectural properties and empirical trade-offs across the evaluated paradigms. ARIMA offers a compact, interpretable structure and remains competitive by MAE. DLinear demonstrates that a small parameter count can suffice to match substantially larger models. LSTM is particularly competitive on ETH, while attention-based models occasionally achieve higher DA without establishing reliable market-direction prediction. Mamba offers selective state processing but shows no consistent MAE leadership. Foundation models enable inference without cryptocurrency-specific parameter updates, although their zero-shot accuracy varies substantially.

These trade-offs concern point accuracy, parameter requirements, and inference resources within the two daily asset series and tested configurations. Parameter counts and asymptotic complexity describe architecture rather than measured deployment cost. Separate latency and memory measurements are discussed below; differences in profiling procedures, batch sizes, and sampling settings limit direct comparisons. Cross-asset, regime-aware, and adapted foundation-model approaches remain unevaluated.

Evidence-Driven Research Gap Analysis

Table 14 separates observed limitations from proposed interventions. Heavy-tailed returns motivate investigation of distributional objectives, while limited directional evidence motivates calibrated direction heads and multi-task evaluation. Rolling cross-asset correlation motivates joint modeling without demonstrating its forecasting benefit. Heterogeneous zero-shot results motivate controlled adaptation studies, and three-fold aggregate evaluation leaves robustness to regime changes unresolved.

These priorities require dedicated experiments under the same temporal controls. The present benchmark identifies questions for further study but does not validate cross-asset conditioning, heavy-tail objectives, domain adaptation, or regime-specific modeling as effective remedies.

Table 13. **Comparative characteristics of forecasting paradigms in the benchmark**

Paradigm	Representative model	Sequence mechanism	Complexity	Main benchmark strength	Main limitation
Statistical	ARIMA	Autoregression	O(N)	Competitive MAE; interpretable	Limited nonlinear modeling
Lightweight DL	DLinear	Linear decomposition	O(N)	High parameter efficiency	Limited adaptive representation
Recurrent	LSTM	Hidden-state recurrence	O(N)	Competitive MAE, especially ETH	Sequential computation
Transformer	PatchTST, iTransformer	Self-attention	$O(N^2)$, or $O(F^2)$/patch-specific	Occasional higher DA	High cost; DA remains near 0.5
Selective SSM	Mamba	Selective state updates	O(N)	Competitive, protocol-dependent	Asset- and metric-dependent
Foundation	Chronos	Large-scale pretraining	High	Model-dependent zero-shot MAE	No domain adaptation tested

***Source*: compiled by the authors**

Table 14. **Evidence-driven research gaps from the literature review and benchmark**

Gap	Evidence status	Benchmark evidence	Next test
G1 – Heavy-tail modeling	Motivated; remedy untested	Large excess kurtosis; MAE/SMAPE rank differences	Compare quantile or Student-t objectives
G2 – Multi-task forecasting	Weak DA observed; remedy untested	DA near 0.5 on normalized returns	Joint return/direction/volatility outputs
G3 – Cross-asset learning	Hypothesis	Strong rolling BTC–ETH correlation; separate asset models	Joint conditioning vs single-asset control
G4 – Foundation adaptation	Hypothesis	Heterogeneous zero-shot performance	Controlled fine-tuning under matched budgets
G5 – Regime-aware evaluation	Limitation observed	Three folds; no regime-stratified metrics	More origins and explicit regime tests

***Source:* compiled by the authors from the literature review and benchmark results**

ISSN 2663-0176 (Print)
ISSN 2663-7731 (Online)

DISCUSSION OF RESULTS

Interpretation of Benchmark Results

The benchmark supports relative model comparisons but shows essentially no absolute MAE gain over zero. BTC ARIMA–LSTM differences are not Holm-supported, and LSTM leads learned-model mean MAE on ETH with many, but not all, supported pairwise advantages. Statistical non-rejection does not establish equivalence. This is a low-signal benchmark rather than evidence of trading value.

Additional capacity does not consistently improve out-of-sample error in these noisy, non-stationary, heavy-tailed series. Since zero is already competitive, small rank differences should not be interpreted as evidence that models recover substantial predictable structure. Financial volatility clustering and anomalous behavior [39] motivate further uncertainty and regime analysis.

A plausible explanation for LSTM's competitiveness is its gated recurrent state, which compresses local temporal dependence within a 60-day window. Flexible attention over patches or correlated feature channels may be more sensitive to noise in a limited daily dataset. Mamba's selective state alone does not guarantee a lower-error representation in a single-asset, single-task setting. These mechanisms are interpretations, not findings established by ablation tests.

Selective State Space Models: Potential and Limitations

Selective State Space Models combine linear sequence complexity with input-dependent state updates, making them attractive for long and non-stationary financial sequences [26], [7]. Recent evidence has also demonstrated the applicability of Mamba to financial forecasting tasks beyond cryptocurrency markets [31]. By compressing historical information into dynamically updated latent states, these models may retain relevant temporal signals without incurring the quadratic cost of full self-attention.

Vanilla Mamba remains competitive in some native-input settings but has no consistent MAE advantage. The aligned control provides negligible best-model skill over zero, limiting claims about additional predictive value from selective memory.

Native-input LSTM has lower ETH mean MAE than Mamba. This does not establish intrinsic unsuitability of state-space modeling: cross-asset conditioning, tail-risk objectives, and regime information require controlled ablations.

Selective memory is therefore an efficient mechanism whose proposed extensions require direct tests. Differences between BTC and ETH alone cannot demonstrate that cross-asset representation learning would improve accuracy.

Model-Dependent Zero-Shot Transfer of Foundation Models

Foundation-model performance varies substantially. TimesFM is comparatively competitive by MAE, Chronos variants are moderate, and Lag-Llama has much larger errors. Low SMAPE for Lag-Llama does not overturn its poor MAE and illustrates the sensitivity of relative-error measures on centered return targets.

At H20, native Lag-Llama signed bias is 0.0221 on BTC and 0.0389 on ETH, compared with MAE 1.0603 and 0.8778. Prediction-to-target standard-deviation ratios are 1.9783 and 1.1394, with correlations −0.0020 and 0.0143. In the aligned control, ratios are 1.9700 and 1.1502 and correlations 0.0041 and 0.0132. These diagnostics suggest excess dispersion rather than a large constant offset. Low SMAPE does not overturn poor MAE; calibration or domain adaptation requires a dedicated test.

Increasing native-input Chronos capacity from 46 million to 710 million parameters gives no consistent MAE benefit. Across foundation families, the 200-million-parameter TimesFM is stronger by MAE than the approximately 2.45-million-parameter Lag-Llama, but parameter count alone does not explain the ranking. Pretraining data, tokenization or patching, and point versus probabilistic output may matter; these are hypotheses rather than causal findings.

The native-input comparison mixes input-set differences with the distinction between pretrained zero-shot models and asset-trained neural models. The univariate control removes additional indicator channels but preserves training and representation differences. Heavy-tailed returns and regime shifts provide plausible domain-gap explanations, but the benchmark does not identify their individual causal effects. Under the aligned control, both Chronos variants still have higher H5 MAE than every supervised neural model and ARIMA (BTC/USDT: 0.493–0.509 versus 0.458–0.478; ETH/USDT: 0.627–0.630 versus 0.568–0.596; Tables 11 and 12), although no model improves on the zero forecast by more than 0.07% in this control. Therefore, the native-input gap is not explained by additional indicator channels alone.

Fine-tuning, domain-specific pretraining, calibration, and exogenous-variable integration

should be evaluated separately with matched adaptation budgets and temporal splits. They are proposed experiments, not demonstrated improvements.

Practical Implications and Future Research Directions

For deployment, model selection should jointly consider point error, information requirements, retraining cost, inference latency, and measured memory usage. ARIMA and DLinear remain useful low-complexity baselines, while LSTM is a strong ETH point-forecasting reference. Table 15 reports inference latency and memory evidence from re-profiling and saved benchmark metrics. Differences in profiling procedures, batch sizes, and forecast sampling limit direct latency and memory comparisons across models.

Weak DA motivates an explicit direction-classification head, directional loss, or multi-task objective combining returns, volatility, and direction probability. Future evaluation should use raw-return direction, class-aware baselines, probability calibration, and transaction costs before asserting trading value.

Further work should directly test quantile or Student-t distributional heads for tail risk; joint BTC-ETH conditioning against single-asset controls; domain adaptation of foundation models; and more rolling-origin folds or explicit bull, bear/crash, sideways, and high-volatility test periods. These are research hypotheses rather than improvements established by this benchmark.

Together, these findings suggest that the next generation of cryptocurrency forecasting systems should combine architectural efficiency with domain-aware objectives, cross-asset information, and explicit uncertainty modeling rather than relying solely on larger sequence models.

CONCLUSIONS

This study contributes an architecture-based taxonomy, a standardized empirical benchmark, and an evidence-driven analysis of research gaps in cryptocurrency forecasting. Ten models were evaluated on daily BTC/USDT and ETH/USDT data using three expanding-window folds and forecasting horizons of 5, 10, and 20 days. Common evaluation periods and metrics enable comparisons across paradigms, while an aligned close-history-only control examines sensitivity to additional input channels where effective inputs change.

No model consistently dominates both assets and all metrics. Best-model MAE skill relative to zero is negligible: native-input gains on BTC reach at most 0.15 %, while ETH skill remains negative. The aligned control yields similarly limited gains. No significant ARIMA–LSTM differences were detected on BTC after Holm adjustment; LSTM achieves the lowest learned-model mean MAE on ETH, although its pairwise advantages are not uniformly significant. TimesFM records the lowest native-input MAE among the evaluated foundation models, and increasing Chronos scale provides no consistent benefit. DA in normalized-return space does not establish raw-price directional accuracy or trading utility. These results suggest that greater model complexity alone offers no reliable accuracy advantage within the evaluated setting.

The evidence is limited to two major daily asset series, three temporal folds, and the tested configurations. It does not establish generalization to smaller assets, intraday data, or additional market regimes. Foundation models were evaluated only zero-shot; probabilistic calibration and trading performance were not assessed. Inference profiling does not capture complete training and deployment costs. Saved point forecasts and targets permit metric recomputation, but exact regeneration of Chronos and Lag-Llama forecasts is not guaranteed because sampling was unseeded and sample paths and generator states were not retained. Limited temporal and stochastic coverage constrains broader conclusions, and the reported pairwise tests establish neither model equivalence nor superiority to zero.

Future work should extend evaluation across assets, frequencies, and regimes, with additional rolling origins and repeated seeded runs. Priorities include heavy-tailed probabilistic objectives, joint forecasting of returns, direction and volatility, cross-asset representation learning, and domain adaptation of foundation models. Selective State Space architectures also warrant investigation for their selective memory mechanisms and linear sequence-processing complexity, although the present Mamba results establish no consistent accuracy advantage. These directions require controlled ablations, regime-specific evaluation, calibration assessment, and analysis of economic performance under transaction costs. The benchmark provides a transparent basis for such investigations without establishing that any proposed extension will improve forecasting.

USE OF ARTIFICIAL INTELLIGENCE

The authors used generative artificial intelligence tools, including Claude (Anthropic) and ChatGPT (OpenAI), to assist with language editing, grammar checking, and improving the clarity and readability of the manuscript. These tools were not

ISSN 2663-0176 (Print)
ISSN 2663-7731 (Online)

used to generate the scientific content, research methodology, experimental design, data analysis, or conclusions. All scientific contributions, interpretations, and final decisions regarding the manuscript were made exclusively by the authors.

ACKNOWLEDGMENTS

This paper is the academic result of research project KHSV2503, funded by Ho Chi Minh City University of Transport.

Table 15. **Computational resource profile of the benchmarked models**

Model	**Parameters**	**Training/adaptation mode**	**Main computational driver**	**Runtime / latency evidence**	**Memory evidence**
ARIMA(5,1,0)	5 AR coefficients	CPU statistical refit per test window	Repeated ARIMA fitting over rolling windows	CPU re-profile on six native-input folds: 50.40 ± 16.41 ms/window	GPU VRAM not applicable; reported CPU RSS delta was below the displayed measurement precision
DLinear	2,440	Supervised training from scratch	Linear trend/residual projections over the input window	CUDA inference re-profile on six native-input folds: 0.04 ± 0.00 ms/window	Peak CUDA inference allocation: 10.7 ± 0.0 MB
LSTM	409,729	Supervised training from scratch	Recurrent encoder-decoder and autoregressive multi-step decoding	CUDA inference re-profile on six native-input folds: 0.19 ± 0.01 ms/window	Peak CUDA inference allocation: 81.0 ± 0.4 MB
PatchTST	613,396	Supervised training from scratch	Self-attention over per-variate patch tokens	CUDA inference re-profile on six native-input folds: 0.07 ± 0.01 ms/window	Peak CUDA inference allocation: 51.2 ± 0.0 MB
iTransformer	605,460	Supervised training from scratch	Self-attention over F=23 variate tokens	CUDA inference re-profile on six native-input folds: 0.07 ± 0.02 ms/window	Peak CUDA inference allocation: 18.2 ± 0.0 MB
Mamba (vanilla)	356,116	Supervised training from scratch	Selective SSM scan with linear sequence scaling	CUDA inference re-profile on six native-input folds: 0.09 ± 0.01 ms/window	Peak CUDA inference allocation: 40.8 ± 0.0 MB
Chronos-T5-Small	46M	Zero-shot inference; no cryptocurrency fine-tuning	Pretrained T5-family decoder inference	Saved six-fold benchmark metrics: 6.19 ± 3.00 ms/window	Representative peak CUDA allocation: 1540.4 MB, batch size 64
Chronos-T5-Large	710M	Zero-shot inference; no cryptocurrency fine-tuning	Larger pretrained T5-family decoder inference	Saved six-fold benchmark metrics: 43.34 ± 2.77 ms/window	Representative peak CUDA allocation: 6419.0 MB, batch size 32
TimesFM	200M	Zero-shot inference; no cryptocurrency fine-tuning	Decoder-only foundation-model inference	Saved six-fold benchmark metrics: 120.38 ± 2.47 ms/window	Representative peak CUDA allocation: 892.2 MB, batch size 1
Lag-Llama	≈2.45M	Zero-shot inference; no cryptocurrency fine-tuning	Probabilistic sample generation from lagged context	Saved six-fold benchmark metrics: 258.86 ± 2.97 ms/window, with 100 samples	Representative peak CUDA allocation: 86.8 MB, batch size 1

Source: **authors' inference profiling and benchmark logs; Lag-Llama parameter count from the official model card**

Conflict of interests: The authors declare no financial, personal, copyright, or other conflicts that could have influenced the study or its results.

Financing: This research was funded by Ho Chi Minh City University of Transport under research project KHSV2503.

Data and code availability: Market data are publicly available from Binance and were accessed through ccxt. Experimental code and benchmark outputs are available from the corresponding author on reasonable request.

REFERENCES

1. Katsiampa, P. "Volatility estimation for Bitcoin: A comparison of GARCH models". *Economics Letters*. 2017; 158: 3–6. DOI: https://doi.org/10.1016/j.econlet.2017.06.023.

2. Chu, J., Chan, S., Nadarajah, S. & Osterrieder, J. "GARCH modelling of cryptocurrencies". *Journal of Risk and Financial Management*. 2017; 10 (4): 17. DOI: https://doi.org/10.3390/jrfm10040017.

3. Gurgul, V., Lessmann, S. & Härdle, W. K. "Deep learning and NLP in cryptocurrency forecasting: integrating financial, blockchain, and social media data". *International Journal of Forecasting*. 2025; 41 (4): 1666–1695, https://www.scopus.com/pages/publications/105000249847.
DOI: https://doi.org/10.1016/j.ijforecast.2025.02.007.

4. Kareem, A. & Aue, A. "Bitcoin forecasting with classical time series models on prices and volatility". *arXiv*. 2025. DOI: https://doi.org/10.48550/arXiv.2511.06224.

5. Nguyen, T. K. T., Antoshchuk, S., Nikolenko, A., Tran, K. T. & Babilunha, O. "Non-stationary time series prediction using one-dimensional convolutional neural network models". *Herald of Advanced Information Technology*. 2020; 3 (1): 362–372. DOI: https://doi.org/10.15276/hait.01.2020.3.

6. Ansari, A. F., Stella, L., Turkmen, C., et al. "Chronos: Learning the language of time series". *Transactions on Machine Learning Research*. 2024, https://www.scopus.com/pages/publications/85217501328. DOI: https://doi.org/10.48550/arXiv.2403.07815.

7. Gu, A. & Dao, T. "Mamba: Linear-time sequence modeling with selective state spaces". In *Proceedings of the First Conference on Language Modeling*. 2024. DOI: https://doi.org/10.48550/arXiv.2312.00752.

8. Nie, Y., Nguyen, N., Sinthong, P. & Kalagnanam, J. "A time series is worth 64 words: Long-term forecasting with transformers". In *Proceedings of the 11th International Conference on Learning Representations*. 2023, https://www.scopus.com/pages/publications/85192521637. DOI: https://doi.org/10.48550/arXiv.2211.14730.

9. Wen, Q., Zhou, T., Zhang, C., Chen, W., Ma, Z., Yan, J. & Sun, L. "Transformers in time series: A survey". In *Proceedings of the 32nd International Joint Conference on Artificial Intelligence*. 2023. p. 6778–6786, https://www.scopus.com/pages/publications/85163733616.
DOI: https://doi.org/10.24963/ijcai.2023/759.

10. Wu, J., Zhang, X., Huang, F., Zhou, H. & Chandra, R. "Review of deep learning models for crypto price prediction: Implementation and evaluation". *Next Research*. 2026; 8: 101513. DOI: https://doi.org/10.1016/j.nexres.2026.101513.

11. Box, G. E. P., Jenkins, G. M., Reinsel, G. C. & Ljung, G. M. "Time series analysis: forecasting and control". 5th ed. *Hoboken, NJ: Wiley*. 2015. DOI: https://doi.org/10.1111/jtsa.12194.

12. Cheikh, N. B., Zaied, Y. B. & Chevallier, J. "Asymmetric volatility in cryptocurrency markets: New evidence from smooth transition GARCH models". *Finance Research Letters*. 2020; 35: 101293. DOI: https://doi.org/10.1016/j.frl.2019.09.008.

13. Kozlov, I. V., Kosenko, V. O. & Kozlova, O. V. "Modeling and forecasting of stock market processes". *Herald of Advanced Information Technology*. 2024; 7 (4): 639–650. DOI: https://doi.org/10.15276/hait.07.2024.7.

14. Hewamalage, H., Bergmeir, C. & Bandara, K. "Recurrent neural networks for time series forecasting: Current status and future directions". *International Journal of Forecasting*. 2021; 37 (1): 388–427. DOI: https://doi.org/10.1016/j.ijforecast.2020.06.008.

15. Guo, Q., Lei, S., Ye, Q. & Zhiyang, F. "MRC-LSTM: A hybrid approach of multi-scale residual CNN and LSTM to predict Bitcoin price". In: *Proceedings of International Joint Conference on Neural Networks*. 2021. p. 1–8, https://www.scopus.com/pages/publications/85116417007. DOI: https://doi.org/10.1109/IJCNN52387.2021.9534453.

16. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł. & Polosukhin, I. "Attention is all you need". In *Advances in Neural Information Processing Systems*. 2017; 30: 5998–6008, https://www.scopus.com/pages/publications/85043317328.

DOI: https://doi.org/10.48550/arXiv.1706.03762.

17. Zhou, H., Zhang, S., Peng, J., Zhang, S., Li, J., Xiong, H. & Zhang, W. “Informer: beyond efficient transformer for long sequence time-series forecasting”. In: *Proceedings of the 35th AAAI Conference on Artificial Intelligence*. 2021; 35 (12): 11106–11115, https://www.scopus.com/pages/publications/85130054909. DOI: https://doi.org/10.1609/aaai.v35i12.17325.

18. Wu, H., Xu, J., Wang, J. & Long, M. “Autoformer: Decomposition transformers with auto-correlation for long-term series forecasting”. In: *Advances in Neural Information Processing Systems*. 2021; 34: 22419–22430, https://www.scopus.com/pages/publications/85130481492.
DOI: https://doi.org/10.48550/arXiv.2106.13008.

19. Zhou, T., Ma, Z., Wen, Q., Wang, X., Sun, L. & Jin, Rg. “FEDformer: Frequency enhanced decomposed transformer for long-term series forecasting”. In: *Proceedings of the 39th International Conference on Machine Learning*. 2022; 162: 27268–27286, https://www.scopus.com/pages/publications/85154558825. DOI: https://doi.org/10.48550/arXiv.2201.12740.

20. Lim, B., Arık, S. Ö., Loeff, N. & Pfister, T. “Temporal fusion transformers for interpretable multi-horizon time series forecasting”. *International Journal of Forecasting* 37, no. 4 (2021): 1748–1764. https://www.scopus.com/pages/publications/85107952273.
DOI: https://doi.org/10.1016/j.ijforecast.2021.03.012.

21. Zhang, Y., & Yan, J. “Crossformer: Transformer utilizing cross-dimension dependency for multivariate time series forecasting”. In: *Proceedings of the 11th International Conference on Learning Representations*. 2023, https://www.scopus.com/pages/publications/85199887879.
DOI: https://doi.org/10.48550/arXiv.2305.18824.

22. Liu, Y., Hu, T., Zhang, H., Wu, H., Wang, S., Ma, L. & Long, M. “iTransformer: Inverted transformers are effective for time series forecasting”. In *Proceedings of the 12th International Conference on Learning Representations*. 2024, https://www.scopus.com/pages/publications/85200562438. DOI: https://doi.org/10.48550/arXiv.2310.06625.

23. Khaniki, M. A. L. & Manthouri, M. “Enhancing price prediction in cryptocurrency using transformer neural network and technical indicators”. *arXiv*. 2024. DOI: https://doi.org/10.48550/arXiv.2403.03606.

24. Mahdi, E., Martin-Barreiro, C. & Cabezas, X. “A novel hybrid approach using an attention-based transformer + GRU model for predicting cryptocurrency prices”. *Mathematics*. 2025; 13 (9): 1484, https://www.scopus.com/pages/publications/105005111432. DOI: https://doi.org/10.3390/math13091484.

25. Patro, B. N. & Agneeswaran, V. S. “SiMBA: Simplified Mamba-based architecture for vision and multivariate time series”. *arXiv*. 2024. DOI: https://doi.org/10.48550/arXiv.2403.15360.

26. Dao, T. & Gu, A. “Transformers are SSMs: Generalized models and efficient algorithms through structured state space dualit”. In: *Proceedings of the 41st International Conference on Machine Learning*. 2024; 235: 10041–10071, https://www.scopus.com/pages/publications/85203797481. DOI: https://doi.org/10.48550/arXiv.2405.21060.

27. Liang, A., Jiang, X., Sun, Y., Shi, X. & Li, K. “Bi-Mamba+: Bidirectional Mamba for time series forecasting”. *arXiv*. 2024. DOI: https://doi.org/10.48550/arXiv.2404.15772.

28. Ma, H., Chen, Y., Zhao, W., Yang, J., Ji, Y., Xu, Xi., Liu, X., Jing, H., Liu, S. & Yang, G. “A Mamba foundation model for time series forecasting”. *arXiv*. 2024. DOI: https://doi.org/10.48550/arXiv.2411.02941.

29. Sepehri, M., Mehradfar, A., Soltanolkotabi, M. & Avestimehr, S. “CryptoMamba: Leveraging state space models for accurate Bitcoin price prediction”. In: *Proceedings of the IEEE International Conference on Blockchain and Cryptocurrency*. 2025. p. 1–3. https://www.scopus.com/pages/publications/105015475354. DOI: https://doi.org/10.1109/ICBC64466.2025.11114565.

30. Wang, Z., Kong, F., Feng, S., Wang, M., Yangm, X., Zhao, H., Wang, D. & Zhang, Y. “Is Mamba effective for time series forecasting?” *Neurocomputing*. 2025; 619: 129178. https://www.scopus.com/pages/publications/85212178664.
DOI: https://doi.org/10.1016/j.neucom.2024.129178.

31. Tran, T. V., Nguyen, T. K. T., Tran, K. T., Pham, T. N. L. & Mai, Q. T. “Enhancing stock price forecasting using the Mamba deep learning model”. *Transportation Science and Technology*. 2025; 14 (4): 77–88. DOI: https://doi.org/10.55228/JTST140408.

32. Das, A., Kong, W., Sen, R. & Zhou, Y. “A decoder-only foundation model for time-series forecasting”. In: *Proceedings of the 41st International Conference on Machine Learning*. 2024 235: 10148–10167, https://www.scopus.com/pages/publications/85203816718.
DOI: https://doi.org/10.48550/arXiv.2310.10688.

33. Gruver, N., Finzi, M., Qiu, S. & Wilson, A. G. “Large language models are zero-shot time series forecasters”. In: *Advances in Neural Information Processing Systems*. 2023; 36, https://www.scopus.com/pages/publications/85187488620. DOI: https://doi.org/10.48550/arXiv.2310.07820.

34. Jin, M., Wang, S., Ma, L., et al. “Time-LLM: Time series forecasting by reprogramming large language models”. In *Proceedings of the 12th International Conference on Learning Representations*. 2024. https://www.scopus.com/pages/publications/85186247296. DOI: https://doi.org/10.48550/arXiv.2310.01728.

35. Woo, G., Liu, C., Kumar, A., Xiong, C., Savarese, S. & Sahoo, D. “Unified training of universal time series forecasting transformers (MOIRAI)”. In: *Proceedings of the 41st International Conference on Machine Learning*. 2024; 235: 53140–53164, https://www.scopus.com/pages/publications/85197071685. DOI: https://doi.org/10.48550/arXiv.2402.02592.

36. Zhou, T., Niu, P., Wang, X., Sun, L. & Jin, R. “One fits all: Power general time series analysis by pretrained LM (GPT4TS)”. In: *Advances in Neural Information Processing Systems*. 2023; 36. https://www.scopus.com/pages/publications/85205555999. DOI: https://doi.org/10.48550/arXiv.2302.11939.

37. Zeng, A., Chen, M., Zhang, L. & Xu, Qiang. “Are transformers effective for time series forecasting?” In: *Proceedings of the 37th AAAI Conference on Artificial Intelligence*. 2023; 37 (9): 11121–11128, https://www.scopus.com/pages/publications/85149166758.
DOI: https://doi.org/10.1609/aaai.v37i9.26317.

38. Rasul, K., Ashok, A., Williams, A. R., et al. “Lag-Llama: Towards foundation models for probabilistic time series forecasting”. *arXiv*. 2023. DOI: https://doi.org/10.48550/arXiv.2310.08278.

39. Khoroshun, A., Ryazantsev V. & Cherpitskyi, O. “Clustering and anomalies of USA stock market volatility index data”. *Advanced Information Systems*. 2023; 7 (2): 12–20. DOI: https://doi.org/10.20998/2522-9052.2023.2.02.

Received 27.09.2026
Received after revision 19.10.2026
Accepted 26.10.2026

DOI: https://doi.org/10.15276/hait.09.2026.35
УДК 004.8:004.89

Глибоке навчання у прогнозуванні цін криптовалют: таксономія, бенчмарк та аналіз дослідницьких прогалин

Нгуєн Тісн Зуй[1)]
ORCID: https://orcid.org/0009-0004-0732-2093; duynguyen.das@gmail.com
Чан Тхе Вінь[1)]
ORCID: https://orcid.org/0000-0002-4241-1065; vinhtt@ut.edu.vn. Scopus Author ID: 56835334700
Нгуєн Тхань Лонг[1)]
ORCID: https://orcid.org/0009-0003-2438-1231, thanhlongnguyen.dev@gmail.com
[1)] Університет транспорту міста Хошимін, вул. Во Оань, 2, Хошимін, 700000, В’єтнам

АНОТАЦІЯ

Актуальність. Дохідності криптовалют характеризуються високою волатильністю, нестаціонарністю та важкими хвостами. Мета. Розробити узгоджену таксономію та еталон порівняння з відтворюваним обчисленням метрик. **Завдання**. Порівняти архітектури й відокремити виміряні обмеження від дослідницьких гіпотез. **Методи**. Десять статистичних, лінійних, рекурентних, увагових, селективних моделей простору станів і попередньо навчених моделей оцінено на щоденних даних Bitcoin/Tether та Ethereum/Tether з дві тисячі вісімнадцятого до дві тисячі двадцять шостого року. Використано три часові розбиття з розширенням навчального вікна та горизонти п’ять, десять і двадцять днів. Одновимірний контроль використовує ті самі календарні рядки після вилучення початкових рядків індикаторів та однакову нормалізацію. Оцінювання охоплює абсолютну похибку, симетричну відсоткову похибку, узгодженість напрямку нормалізованої дохідності й парні тести з корекцією часової залежності та множинних порівнянь. **Наукова новизна**. Поєднано таксономію, узгоджений контроль вхідної інформації та нульовий прогноз. **Практична значущість**. Результати допомагають оцінювати точність і обчислювальні витрати за слабкого прогнозного сигналу. **Результати**. Найкраще покращення абсолютної похибки відносно нульового прогнозу є практично нульовим. Для Bitcoin відмінність

авторегресійної та рекурентної моделей не підтверджена скоригованими тестами. Для Ethereum рекурентна модель має найменшу середню абсолютну похибку серед навчених моделей, але підтверджено не всі парні переваги. Узгодженість напрямку близька до нуля цілих п’яти десятих. TimesFM є найсильнішою фундаментальною моделлю за абсолютною похибкою у протоколі з власними входами; збільшення Chronos не дає стабільної переваги. Lag-Llama має надмірну дисперсію прогнозів, тому низька симетрична відсоткова похибка не означає вищої точності. **Висновки**. Еталон зі слабким сигналом підтримує порівняння в межах протоколу, а не твердження про прибуткове прогнозування. Узагальнення обмежене двома активами, трьома розбиттями та оцінюванням фундаментальних моделей без адаптації.

Ключові слова: прогнозування криптовалют; глибоке навчання; часові ряди; трансформери; Mamba; фундаментальні моделі

ABOUT THE AUTHORS



Nguyen Tien Duy - Student of Data Science major, Institute of Information Technology & Electrical and Electronics Engineering, Ho Chi Minh City University of Transport, 2, Vo Oanh Str., Ho Chi Minh City, 700000, Vietnam
ORCID: https://orcid.org/0009-0004-0732-2093; duynguyen.das@gmail.com
Research field: Data science, Deep Learning, Computer Vision, Natural Language processing

Нгуєн Тієн Зуй - студент спеціальності «Наука про дані», Інститут інформаційних технологій, електротехніки та електронної інженерії, Університету транспорту міста Хошимін, вул. Во Оань, 2, Хошимін, 700000, В’єтнам
Галузь досліджень: Наука про дані, глибоке навчання, комп'ютерний зір, обробка природної мови



Tran The Vinh - Doctor of Philosophy, Program Director of Data Science, Ho Chi Minh City University of Transport, 2, Vo Oanh Str., Ho Chi Minh City, 700000, Vietnam
ORCID: https://orcid.org/0000-0002-4241-1065; vinhtt@ut.edu.vn. Scopus Author ID: 56835334700
Research field: Data Science, Cybersecurity, Blockchain Technologies, Deep Learning and Systems Architecture

Чан Тхе Вінь - доктор філософії, директор програми з науки про дані, Університет транспорту міста Хошимін, вул. Во Оань, 2, Хошимін, 700000, В’єтнам
Галузь досліджень: Наука про дані, кібербезпека, технології блокчейн, глибоке навчання та системна архітектура



Nguyen Thanh Long – Student of Data Science major, Institute of Information Technology & Electrical and Electronics Engineering, Ho Chi Minh City University of Transport, 2, Vo Oanh Str., Ho Chi Minh City, 700000, Vietnam
ORCID: https://orcid.org/0009-0003-2438-1231, thanhlongnguyen.dev@gmail.com
Research field: Data science, Deep Learning, Natural Language processing, Systems Architecture

Нгуєн Тхань Лонг - студент спеціальності «Наука про дані», Інститут інформаційних технологій, електротехніки та електронної інженерії, Університету транспорту міста Хошимін, вул. Во Оань, 2, Хошимін, 700000, В’єтнам
Галузь досліджень: Наука про дані, Глибоке навчання, Обробка природної мови, Архітектура систем