ISSN: 3080-9843 (Print), ISSN: 3080-9851 (Online)

How Much Do Sentiment and On-Chain Data Add? A HAR-Anchored Multimodal Mamba for Bitcoin Volatility Forecasting

Lily Jenkins [1, *], Yuna Rao [2]

[1] School of Computing and Information, University of Pittsburgh, USA

[2] Oracle Corporation, USA

[*] Corresponding Author: Lil.jenkins28@outlook.com

Abstract

Sentiment indices and blockchain activity are widely proposed as predictors of cryptocurrency risk, yet their value for volatility forecasting beyond strong econometric benchmarks is rarely tested under strict out-of-sample conditions. We present SOC-Mamba, a multimodal selective state space framework that learns a residual on top of a heterogeneous autoregressive (HAR) forecast and issues heteroscedastic 1-, 7- and 30-day forecasts from market, Fear & Greed sentiment and Coin Metrics on-chain streams. Two fusion strategies are compared: early fusion (EF) and a cross-modal selective scan (XS), which extends the volatility-aware selection of CryptoMamba-SSM so that a reliability-gated auxiliary context also conditions the step size and projections of the Mamba encoder. On a public-data benchmark of daily Bitcoin realized range-based volatility from 4-hour bars (2020–2026) with explicit publication lags, 13 forecasters are compared over five expanding-window folds, with capacity and training length chosen on validation data only. Without the anchor, deep sequence models have 9.3–53.5% higher MSE than HAR, which identifies the anchor as the key ingredient. Both anchored variants join HAR in the 90% model confidence set at every horizon; SOC-Mamba-EF attains the lowest 1-day MSE and reduces MSE by 10.7–23.7% during the FTX collapse, and SOC-Mamba-XS improves significantly on HAR in 1-day QLIKE. Econometric anchoring makes deep multimodal forecasters competitive with strong econometric benchmarks, and the auxiliary sentiment and on-chain data deliver their largest gains during stress episodes.

Keywords

Bitcoin, realized volatility, state space models, Mamba, multimodal fusion, investor sentiment, on-chain data, HAR model, forecast evaluation.

1. Introduction

Volatility is the main input to risk limits, margining and option pricing, and in cryptocurrency markets it is both high and unstable [1]. Bitcoin's daily volatility can triple within weeks, and its average level has drifted downward as the market has matured, which makes any forecaster estimated on past data vulnerable to regime change. Simple econometric models are hard to beat in this setting. GARCH-type models capture clustering and asymmetry [2], and heterogeneous autoregressive (HAR) models of realized volatility [3] have outperformed them on Bitcoin when intraday data are available [4].

Two further information sources are observable every day. Attention and sentiment measures co-move with Bitcoin prices and trading activity [5, 6], and blockchain data such as active addresses, fees, hash rate and exchange flows describe network usage and the supply that is ready to be sold [7]. These signals are attractive inputs for deep forecasters, but they are noisy, partly redundant with prices and published with delays, and their relation to volatility may

ISSN: 3080-9843 (Print), ISSN: 3080-9851 (Online)

change between regimes. Whether they improve volatility forecasts once a realized-volatility benchmark is included, and once every modeling choice is made without access to the test period, remains an open empirical question.

Selective state space models are a natural backbone for daily financial sequences. Mamba [8] makes the discretization step and the projections of a structured state space model (SSM) [9] depend on the current input, so the model decides at each step how much of its state to keep, at a cost linear in sequence length. Mamba-based models have been applied to Bitcoin price prediction [10] and to generic time series forecasting [11], and the selection mechanism has recently been adapted to cryptocurrency volatility (Section II-C). These studies use market data only and seldom compare against HAR, the standard benchmark for realized volatility.

This paper asks how sentiment and on-chain information should enter a selective SSM for volatility forecasting, and how much they add under a strict protocol. We propose SOC-Mamba (Sentiment, On-Chain and market Mamba), a framework built on three ideas. First, the network forecasts a residual on top of a HAR anchor and starts training from the HAR solution, so it departs from the benchmark only where the data support it. Second, it produces heteroscedastic forecasts for 1-, 7- and 30-day horizons with a β-weighted likelihood. Third, auxiliary modalities enter either by early fusion (SOC-Mamba-EF) or through a cross-modal selective scan (SOC-Mamba-XS), in which a reliability-gated summary of sentiment and on-chain encoders conditions the step size Δt and the projections Bt, Ct of the market SSM. Because Δt sets the decay rate of each state channel, the XS design lets exogenous shocks shorten the memory of the volatility state instead of shifting the forecast linearly.

The contributions are as follows. (i) We formulate the cross-modal selective scan with reliability gating, availability masks and modality dropout, together with a HAR-anchored heteroscedastic multi-horizon head. (ii) We assemble a daily multimodal Bitcoin benchmark from public sources (2020–2026) with explicit publication lags and realized range-based targets, and an evaluation protocol in which capacity, regularization and training length are chosen on validation data only. (iii) We compare 13 forecasters with Diebold–Mariano tests and model confidence sets and analyze stress episodes, feed outages and interval calibration. (iv) We show that HAR anchoring is decisive: unanchored deep sequence models trail HAR, whereas the anchored variants reach the HAR level, enter the model confidence set and improve on it during stress episodes and in 1-day QLIKE. Auxiliary data help the early-fusion variant only in stress periods and give the cross-modal variant no clear incremental gain.

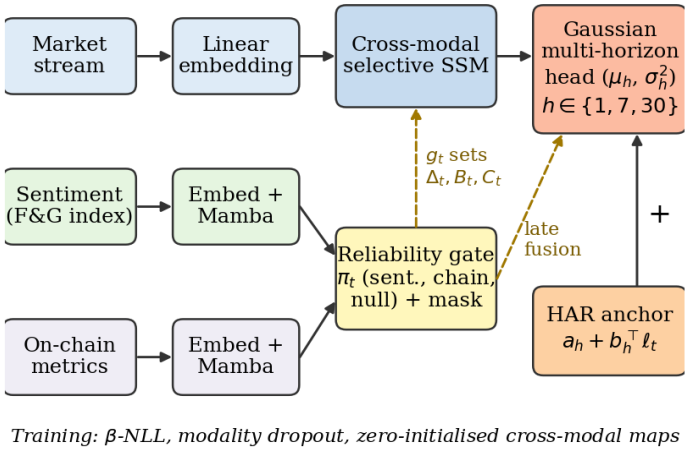


Figure 1. SOC-Mamba with the cross-modal selective scan (XS). The fused auxiliary context gt conditions the selection parameters of the market SSM and also enters the head through a zero-initialized late-fusion path. In the early-fusion variant (EF), all inputs are concatenated before a single Mamba encoder; the HAR anchor, head and loss are identical.

ISSN: 3080-9843 (Print), ISSN: 3080-9851 (Online)

2. Related Work

2.1. Volatility Forecasting for Cryptocurrencies

GARCH models [12, 13] remain the default when only daily returns are used; for Bitcoin, Katsiampa [2] found that specifications with separate short- and long-run volatility components fit best. With intraday data, realized measures allow volatility to be modeled directly. The HAR model [3] approximates long memory with daily, weekly and monthly components, and Bergsli et al. [4] showed that HAR models on realized variance outperform GARCH models for Bitcoin, especially at short horizons. Range-based estimators [14] and the realized range [15] use highs and lows and are more efficient than squared returns when sampling is coarse. Machine learning can beat HAR when data are plentiful: Christensen et al. [16] report gains for Dow Jones constituents, partly from additional predictors and partly from better persistence. With one asset and a few years of daily data the room for such gains is much smaller, which motivates the anchored design used here. Decision-focused demand-forecasting research illustrates how risk-aware objectives can assess predictive models according to the quality of downstream constrained decisions [31].

2.2. Sentiment and On-Chain Information

Kristoufek [5] documented a two-way relation between Bitcoin prices and search and Wikipedia attention, and Shen et al. [6] found that tweet volume predicts next-day trading volume and realized volatility rather than returns. Blockchain variables have been used to model the Bitcoin price process, for example in the Bayesian neural network study of Jang and Lee [7]. Most of this evidence is in-sample or concerns returns; less is known about the incremental out-of-sample value of such data for multi-horizon realized volatility forecasts.

2.3. State Space Models in Finance

Structured SSMs [9] parameterize long convolution kernels through linear time-invariant state equations, and Mamba [8] makes the parameters input-dependent. S-Mamba [11] matched Transformer-based forecasters at lower cost on public benchmarks, and CryptoMamba [10] applied Mamba to daily Bitcoin price prediction and trading. Closest to our work, CryptoMamba-SSM [17] targets cryptocurrency volatility directly. It conditions B_t, C_t and Δ_t on the previous volatility estimate and on realized-volatility indicators, uses separate input matrices for positive and negative drive, mixes regime-specific state matrices and trains a multi-horizon head with a heteroscedastic Gaussian loss, with clear gains over recurrent and attention baselines on Bitcoin data. Its central insight, that the selection mechanism is the right place to inject financial state information, is the starting point of the cross-modal selective scan proposed here, which extends the conditioning signal from market indicators to other modalities. GAN-enhanced implied-volatility surface reconstruction offers another financial-market example of representing nonlinear risk structure from market-derived data [32].

2.4. Multimodal Fusion and Missing Modalities

Multimodal learning distinguishes early, late and intermediate fusion [18]. Gated fusion weights modalities by their estimated usefulness, and ModDrop [19] removes entire modalities at random during training so that a model degrades gracefully when inputs are missing. Both ideas are relevant for market data, where sentiment indices and blockchain analytics come from different providers with different delays and occasional gaps.

ISSN: 3080-9843 (Print), ISSN: 3080-9851 (Online)

3. Data and Problem Formulation

3.1. Data Sources and Timing

We combine three public sources. Prices are 4-hour BTC/USD open, high, low and close values from the CoinGecko aggregate feed between 25 March 2020 and 20 October 2026, with no missing bars. Sentiment is the daily Crypto Fear & Greed Index [20], a 0–100 composite of volatility, momentum, social media activity, dominance and search trends; one missing day (26 October 2024) was forward-filled, and the series agrees with an independently archived copy on all 1,804 overlapping days. On-chain data are the Coin Metrics community metrics for Bitcoin [21], which end on 23 May 2026. After warm-up periods for rolling features, the sample has 2,221 daily forecast origins from 25 April 2020 to 24 May 2026.

A forecast is issued at the end of UTC day t. Price features use the six bars that open on day t. The Fear & Greed value used at origin t is the one dated t, published at the start of that day, and all Coin Metrics variables, including reported spot volume, are lagged by one day to allow for publication delays. Every input is therefore available at the time the forecast is made.

3.2. Volatility Measurement and Targets

Range-based estimation is attractive when intraday sampling is coarse, and CryptoMamba-SSM [17] likewise builds its volatility targets from range-based estimators. We apply the Garman–Klass estimator [14] to each 4-hour bar and sum over the day, in the spirit of the realized range [15]:

$$RV_t = \sum_{j=1}^{6} \left[\tfrac{1}{2} \ln^2(H_{tj}/L_{tj}) - (2\ln 2 - 1) \ln^2(C_{tj}/O_{tj}) \right] \quad (1)$$

Where H_{tj}, L_{tj}, O_{tj} and C_{tj} are the high, low, open and close of bar j. Because the market trades continuously and each bar opens at the previous close, no overnight term is needed. The target for horizon h is the log volatility of the average future variance,

$$y_t(h) = \tfrac{1}{2} \ln\left(h^{-1} \sum_{i=1}^{h} RV_{t+i} \right), \quad h \in \{1, 7, 30\} \quad (2)$$

As a robustness proxy we also compute realized variance from squared 4-hour log returns; its daily log volatility has a correlation of 0.84 with the range-based measure. The HAR components at origin t are $\ell_{td} = \tfrac{1}{2} \ln RV_t$ and the log volatilities ℓ_{tw} and ℓ_{tm} of the 7- and 30-day average variances; windows are measured in calendar days because the market never closes. The HAR benchmark is the regression estimated separately for each horizon.

$$y_t(h) = \beta_0 + \beta_d \ell_{td} + \beta_w \ell_{tw} + \beta_m \ell_{tm} + \varepsilon_t(h) \quad (3)$$

3.3. Input Modalities

Table I lists the 26 inputs. The market stream contains the three HAR components, the log realized variance of 4-hour returns, the daily return and its absolute value, the downside share of realized semivariance, 7- and 30-day momentum and abnormal spot volume. The sentiment stream contains the index level, its 1- and 7-day changes, a 30-day z-score and flags for extreme fear and extreme greed. The on-chain stream contains abnormal activity, defined as the deviation of the log value from its 30-day mean, for active addresses, transactions, fees, hash rate and exchange inflows and outflows, together with the MVRV ratio and its weekly change, net exchange flow as a share of exchange-held supply and the 30-day change of exchange-held supply. Inputs are standardized with the training-window statistics of each fold and clipped at ±5.

ISSN: 3080-9843 (Print), ISSN: 3080-9851 (Online)

Table I. Input modalities (26 daily variables)

Stream	Variables
Market (10)	HAR components ℓd, ℓw, ℓm; log RV of 4-h returns; return; |return|; downside semivariance share; 7- and 30-day momentum; abnormal spot volume (lag 1)
Sentiment (6)	Fear & Greed level; 1- and 7-day changes; 30-day z-score; extreme-fear (≤25) and extreme-greed (≥75) flags
On-chain (10, lag 1)	Abnormal active addresses, transactions, fees, hash rate, exchange inflow and outflow; MVRV and its weekly change; net exchange flow (% of exchange supply); 30-day change of exchange supply

Fig. 2 shows the main series. Volatility fell after 2022 while sentiment and MVRV went through two full cycles, so any fixed relation between these variables and the volatility level is fragile.

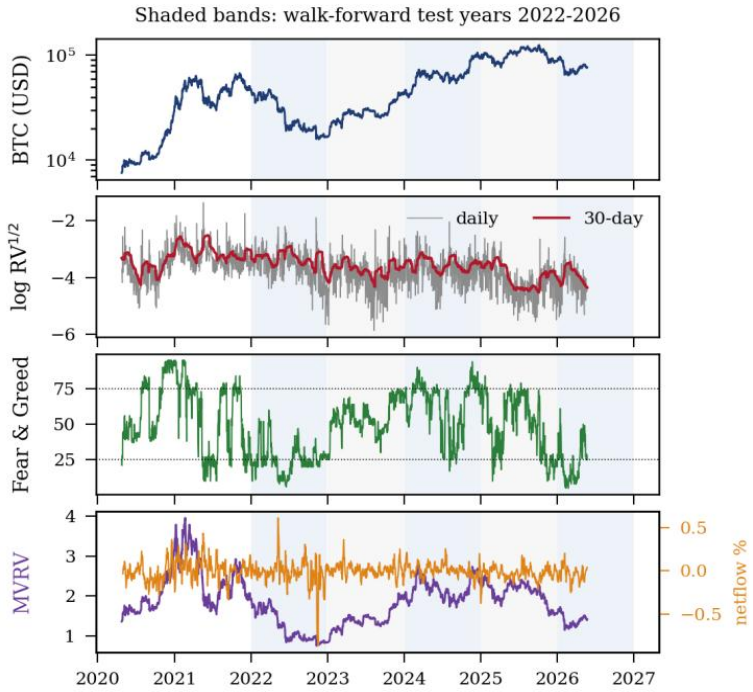


Figure 2. BTC price, daily and 30-day realized log volatility, Fear & Greed Index, MVRV ratio and 7-day net exchange flow. Shaded bands mark the five walk-forward test periods.

3.4. Walk-Forward Protocol

We use five expanding-window folds with test periods 2022, 2023, 2024, 2025 and 2026 (January to 24 May). For test year Y, the 120 days before 1 January of Y form the validation period; only validation origins whose 30-day target window ends before the test year are kept, and training origins end a further 31 days earlier, so no target used for fitting or model selection overlaps a later period. Training sets grow from 407 to 1,868 origins, each validation set has 90 origins, and the pooled test set has 1,605 origins. It covers the Terra/LUNA and FTX collapses of 2022, the spot-ETF year 2024 and the liquidation cascade of October 2025.

4. Methodology

4.1. Overview

Fig. 1 summarizes SOC-Mamba. Each stream m ∈ {M, S, O} (market, sentiment, on-chain) is embedded linearly. In the XS variant, the sentiment and on-chain streams pass through lightweight Mamba encoders whose outputs are fused by a reliability gate into a context sequence gt, and the market stream passes through a Mamba encoder whose selection

ISSN: 3080-9843 (Print), ISSN: 3080-9851 (Online)

parameters depend on gt and on the realized-volatility state. In the EF variant, all 26 inputs are concatenated before a single Mamba encoder. In both cases the representation at the forecast origin feeds a heteroscedastic multi-horizon head that adds a residual to a HAR anchor.

4.2. Selective State Space Encoder

A Mamba block [8] applies an input projection and a causal depthwise convolution, after which each channel of the signal $u_t \in \mathbb{R}^P$ drives a diagonal state $h_t \in \mathbb{R}^{P\times N}$:

$$h_t = \exp(\Delta_t A) \odot h_{t-1} + \Delta_t B_t u_t, \quad y_t = C_t^\top h_t + D \odot u_t \tag{4}$$

With zero-order-hold discretization of a fixed diagonal $A = -\exp(A_{log})$ and input-dependent parameters $\Delta_t = \mathrm{softplus}(W_\Delta u_t + b_\Delta)$, $B_t = W_B u_t$ and $C_t = W_C u_t$. The output is gated by a SiLU branch and projected back with a pre-norm residual connection. For a sequence of length L the cost is O(LPN).

4.3. Reliability-Gated Fusion

Let e_t^S and e_t^O be the outputs of the auxiliary encoders, e^M the market embedding and $a \in \{0, 1\}^2$ the availability flags. The gate has three outputs: sentiment, on-chain and a null slot. Logits of unavailable streams are set to $-\infty$, and the null slot lets the model ignore both streams. Because π_t varies over time, the weighting can differ between calm and turbulent periods.

$$\pi_t = \mathrm{softmax}(W_\pi [e^M; e_t^S; e_t^O; a] + b_\pi), \quad g_t = \pi_t^S e_t^S + \pi_t^O e_t^O \tag{5}$$

4.4. Cross-Modal Selective Scan

In the XS variant the selection of the market encoder becomes

$$\Delta_t = \mathrm{softplus}(W_\Delta u_t + U_\Delta c_t + b_\Delta) \tag{6}$$

$$B_t = W_B u_t + U_B c_t, \quad C_t = W_C u_t + U_C c_t \tag{7}$$

Where $c_t = [g_t; v_t]$ stacks the fused context and the standardized daily log realized volatility v_t. Conditioning on v_t follows the volatility-aware selection of CryptoMamba-SSM [17]; the new element is that information from other modalities reaches the same parameters. We also adopt its asymmetric update, which splits the pre-activation into positive and negative parts with separate input matrices,

$$h_t = \exp(\Delta_t A) \odot h_{t-1} + \Delta_t \odot (B_t^+ u_t^+ + B_t^- u_t^-) \tag{8}$$

For channel p and state n the per-step decay factor is $\exp(-\Delta_{tp}|A_{pn}|)$, so the effective memory length is

$$\tau_{tpn} = 1 / (\Delta_{tp} |A_{pn}|) \tag{9}$$

A context that raises Δ_t shortens the memory and lets the newest observation dominate the state; a context that lowers it extends the memory. Sentiment and on-chain shocks therefore act on how fast the volatility state forgets, rather than adding a linear term to the forecast. The cross-modal maps U_Δ, U_B and U_C are initialized at zero, so training starts from a market-only model.

ISSN: 3080-9843 (Print), ISSN: 3080-9851 (Online)

4.5. HAR-Anchored Heteroscedastic Head

Let zt be the normalized encoder output at the forecast origin, plus a zero-initialized projection of gt in the XS variant. For each horizon h,

$$\mu t(h) = ah + bh^{\top} \ell t + wh^{\top} zt, \quad \ln \sigma t2(h) = qh^{\top} zt + kh \tag{10}$$

Where $\ell t = (\ell td, \ell tw, \ell tm)$. The anchor (ah, bh) is initialized with the HAR least-squares solution on the training window and wh with zeros, so the untrained network reproduces HAR exactly. With eth = yt(h) − μt(h), the loss is the β-NLL of Seitzer et al. [22],

$$\mathcal{L} = \text{mean}_{t,h} \lfloor \sigma th2\beta \rfloor (\tfrac{1}{2} \ln \sigma th2 + eth2 / 2\sigma th2) \tag{11}$$

Where $\lfloor \cdot \rfloor$ stops the gradient, β = 0.5 and the mean runs over all supervised positions and horizons. With β = 0 the loss reduces to the Gaussian negative log-likelihood used by CryptoMamba-SSM; the weighting keeps the network from lowering the loss by inflating the predicted variance of hard observations [22].

4.6. Training and Model Selection

All encoders are causal, so we train sequence-to-sequence on chunks of 120 consecutive days and compute the loss at every position with at least 60 days of context; at evaluation time, forecasts use 90-day windows ending at the origin. This reuses each origin many times per epoch and is about 30 times cheaper than training on sliding windows. In the XS variant, each auxiliary stream of a chunk is zero-filled and flagged as unavailable with probability 0.15 (modality dropout [19]). We use AdamW [23] with learning rate 10^{-3}, weight decay 10^{-4}, batches of eight chunks and gradient clipping at 1.0.

Early stopping on 90-day validation windows proved unreliable in preliminary runs, where it usually stopped within a few epochs. We therefore train every run for 60 epochs, keep snapshots at 13 epochs between 1 and 60, and choose one checkpoint per model configuration with the pooled criterion over K = 5 folds and S = 3 seeds. Dividing by the HAR loss of each validation window keeps the fold that contains the FTX collapse from dominating the choice. Capacity was selected in the same way (Section V-C). No test observation is used for any choice, and every neural forecast is the average of three seeds.

$$c^* = \text{argmin}_c (KS)^{-1} \Sigma_{k,s} MSEksval(c) / MSEkval(HAR) \tag{12}$$

5. Experimental Setup

5.1. Baselines

We compare 13 forecasters. The econometric group contains a random walk on the volatility level (the trailing h-day realized log volatility); HAR [3], estimated by OLS on training and validation origins; HAR-X, which adds the 16 auxiliary variables with a ridge penalty chosen on validation; and GARCH(1, 1) [12] and GJR-GARCH(1, 1) [13] with Student-t innovations on daily returns, whose average variance forecasts are mapped to the target scale by a Mincer–Zarnowitz regression fitted on the training origins. LightGBM [24] uses all 26 inputs with early stopping on validation. The deep sequence models are an LSTM [25] and a causal Transformer encoder [26] with early fusion, a market-only Mamba, an early-fusion Mamba (Mamba-EF) and CM-SSM core†, our re-implementation of the selective core of CryptoMamba-SSM [17] on the market stream.

ISSN: 3080-9843 (Print), ISSN: 3080-9851 (Online)

CM-SSM core† contains volatility-aware selection conditioned on the three HAR components and the absolute return, the asymmetric update, a regime-aware state matrix At = Abase + Σk βtk Ak with K = 4 regimes (mixed in log space to keep At negative) and the heteroscedastic multi-horizon loss. The core retains the selective components that apply to daily inputs, while the multi-scale feature extraction and volatility-aware attention of the original model target high-frequency data. None of the deep baselines uses the HAR anchor. SOC-Mamba-EF and SOC-Mamba-XS share the anchor, head, loss and training procedure and differ only in how the auxiliary streams enter; the EF variant was specified in advance as a controlled comparison of the fusion mechanism.

5.2. Metrics and Tests

Point forecasts are evaluated by the mean squared error (MSE) of the log volatility, the training objective of most models, and by QLIKE on the variance scale,

$$\text{QLIKE} = V/F - \ln(V/F) - 1, \quad V = e^{2y}, \quad F = e^{2\hat{y}} \tag{13}$$

Which belongs to the class of loss functions that rank forecasts consistently when the volatility proxy is noisy but unbiased [27]. Equal accuracy is tested with the Diebold–Mariano (DM) test [28] with the small-sample correction of Harvey et al. [29] and a Bartlett HAC variance with h − 1 lags. The 90% model confidence set (MCS) [30] under MSE loss is computed with a block bootstrap and 2,000 replications. Prediction intervals are judged by the coverage and width of central 90% Gaussian intervals in log-volatility space.

5.3. Implementation and Capacity Selection

All models were implemented in JAX and trained on a single CPU core. Three capacity settings of SOC-Mamba-XS were compared with criterion (12). The initial setting (width 32, two blocks, eight states) scored 1.0007 with three seeds because the network over-fitted after the first epoch. A compact setting (width 16, one block, four states, auxiliary width 8) scored 0.9899 with weight decay 10−4 and dropout 0.1 and 0.9899 with weight decay 0.05 and dropout 0.2 (one seed each). As the two compact settings were tied, we adopted the one with lighter regularization and used it for all neural models. SOC-Mamba-XS then has 5,539 parameters, SOC-Mamba-EF 3,102, CM-SSM core† 3,746, the LSTM 7,651 and the Transformer 2,739. One run of SOC-Mamba-XS (one fold, one seed, 60 epochs) takes about 7 s, and all 225 neural runs take about 20 minutes. The selected checkpoints are epoch 2 for SOC-Mamba-XS, 10 for SOC-Mamba-EF and between 25 and 40 for the unanchored networks.

6. Results

6.1. Overall Accuracy

Table II reports pooled out-of-sample accuracy. HAR remains the reference: it is in the 90% MCS at every horizon, and neither the GARCH models nor LightGBM come close to it in MSE. Adding the auxiliary variables linearly (HAR-X) raises the MSE by 6.7%, 10.2% and 14.2% at the three horizons. All unanchored deep models are clearly worse than HAR, with MSE between 9.3% and 53.5% higher. CM-SSM core† improves on the LSTM at every horizon but stays at the level of the other Mamba baselines, and the Transformer is the strongest member of the group at the 1- and 30-day horizons. None of these models is in the MCS at h = 1 or h = 7, and only the Transformer is at h = 30.

ISSN: 3080-9843 (Print), ISSN: 3080-9851 (Online)

Table II. Out-of-sample accuracy, 2022-01-01 to 2026-05-24 (1,605 origins)

	MSE (log volatility)			QLIKE (variance)		
Model	h=1	h=7	h=30	h=1	h=7	h=30
Econometric						
Random walk	0.2554	0.1165‡	0.1079‡	0.7346	0.3274	0.2393
HAR	0.1830‡	0.1003‡	0.1117‡	0.4586	0.2294	0.1785
HAR-X (ridge)	0.1952	0.1106‡	0.1277‡	0.4674	0.2361	0.1948
GARCH(1, 1)-t	0.2335	0.1364	0.1486	0.5779	0.2959	0.2231
GJR-GARCH-t	0.2357	0.1357	0.1451	0.6111	0.3085	0.2226
Machine learning						
LightGBM	0.2037	0.1885	0.2108	0.4310	0.2883	0.2868
Deep sequence models (no anchor)						
LSTM	0.2310	0.1541	0.1649	0.4747	0.2876	0.2497
Transformer	0.2000	0.1327	0.1286‡	0.4788	0.3049	0.2241
Mamba (market)	0.2060	0.1279	0.1408	0.4750	0.2793	0.2261
CM-SSM core†	0.2088	0.1310	0.1400	0.4697	0.2835	0.2185
Mamba-EF	0.2086	0.1354	0.1593	0.4643	0.2702	0.2530
HAR-anchored (proposed)						
SOC-Mamba-EF	0.1818‡	0.1014‡	0.1108‡	0.4476	0.2360	0.1837
SOC-Mamba-XS	0.1830‡	0.1003‡	0.1129‡	0.4521	0.2318	0.1813

Best value per column in bold. ‡ member of the 90% model confidence set (MSE loss, 13 models). † re-implemented selective core of CryptoMamba-SSM (Section V-A). Neural forecasts are averages over three seeds.

Both anchored variants join HAR in the MCS at every horizon. SOC-Mamba-EF has the lowest 1-day MSE of all models (0.1818 against 0.1830 for HAR) and the lowest 30-day MSE apart from the random walk, which profits from the fall in volatility after 2022. SOC-Mamba-XS is almost indistinguishable from HAR in MSE (0.1830, 0.1003 and 0.1129) but lowers the 1-day QLIKE from 0.4586 to 0.4521. Directional accuracy tells the same story: HAR, SOC-Mamba-EF and SOC-Mamba-XS predict the direction of the 1-day volatility change correctly at 69.6%, 70.3% and 68.9% of the origins.

6.2. Statistical Significance

Table III reports DM statistics of the two anchored variants against the other models. Both are significantly more accurate than the GARCH models, LightGBM and the unanchored deep models at almost all horizons; the exception is the Transformer at h = 30 (p = 0.200 for XS). Against HAR, neither variant differs significantly under MSE (SOC-Mamba-EF: p = 0.335, = 0.608 and = 0.796). Under QLIKE, SOC-Mamba-XS is significantly better than HAR at h = 1 (DM = -3.03, p = 0.002) but worse at h = 30 (DM = 2.56, p = 0.011), and SOC-Mamba-EF has a lower but not significantly different 1-day QLIKE (p = 0.071). The two variants are not significantly different from each other.

ISSN: 3080-9843 (Print), ISSN: 3080-9851 (Online)

Table III. Diebold–Mariano tests (MSE loss) of the anchored variants

	SOC-Mamba-XS vs.			SOC-Mamba-EF vs.		
Model	h=1	h=7	h=30	h=1	h=7	h=30
Random walk	−12.1***	−2.3**	0.3	−12.1***	−2.0**	0.2
HAR	−0.0	0.0	1.6	−1.0	0.5	−0.3
HAR-X (ridge)	−4.0***	−2.1**	−1.3	−4.5***	−2.1**	−1.7*
GARCH(1, 1)-t	−8.9***	−5.3***	−3.5***	−8.8***	−5.0***	−3.3***
GJR-GARCH-t	−9.2***	−5.3***	−3.3***	−9.3***	−5.2***	−3.3***
LightGBM	−5.1***	−5.9***	−4.4***	−5.4***	−5.9***	−4.5***
LSTM	−9.0***	−5.3***	−3.0***	−9.4***	−5.5***	−3.3***
Transformer	−4.6***	−4.5***	−1.3	−5.3***	−5.0***	−1.8*
Mamba (market)	−5.6***	−5.1***	−3.4***	−5.9***	−5.3***	−3.6***
CM-SSM core†	−5.8***	−5.2***	−3.1***	−6.1***	−5.5***	−3.2***
Mamba-EF	−6.0***	−4.8***	−3.8***	−6.6***	−5.1***	−4.6***
Other anchored variant	1.2	−0.6	0.7	−1.2	0.6	−0.7

HLN-corrected statistics; negative values favor the anchored variant. ***, **, *: two-sided $p < 0.01$, 0.05, 0.10.

6.3. Results by Year and in Stress Episodes

Fig. 3 breaks the MSE ratio to HAR down by test year. SOC-Mamba-XS stays within 2.6% of HAR in every year and horizon and SOC-Mamba-EF within 14.7%, whereas the random walk loses heavily in most years but gains strongly at h = 30 in 2025, when volatility stayed far below its training-period mean. SOC-Mamba-EF helps most at the 30-day horizon in 2022 and 2026, with MSE ratios of 0.91 and 0.91, and loses in 2024 (1.15).

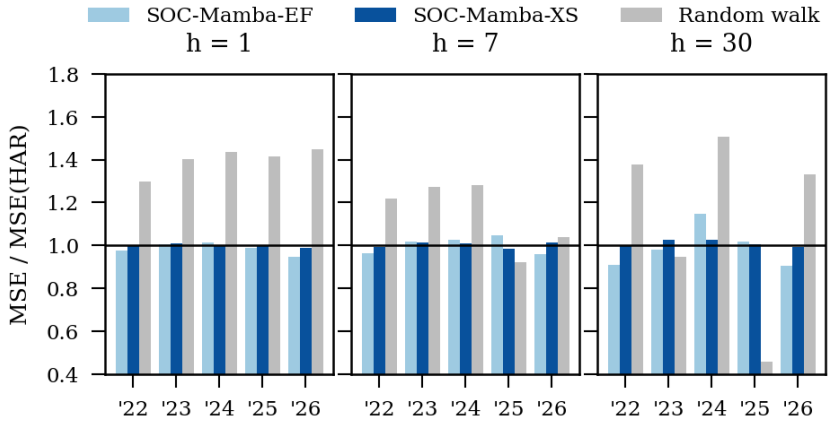


Figure 3. MSE relative to HAR by test year and horizon (values below 1 favor the model).

Table IV isolates four stress episodes. During the FTX collapse SOC-Mamba-EF reduced the MSE relative to HAR by 10.7% at h = 1 and 23.7% at h = 7, and during the August 2024 sell-off by 3.0% and 9.8%. The pattern is not uniform. In the LUNA/3AC episode HAR and GARCH were more accurate at h = 7, and in October 2025 CM-SSM core† had the lowest error at both horizons. Each window contains only 21 to 45 origins, so these results indicate where the multimodal network adds value rather than establishing it statistically.

ISSN: 3080-9843 (Print), ISSN: 3080-9851 (Online)

Table IV. MSE during market stress episodes

	HAR	GARCH	CM-SSM†	SOC-EF	SOC-XS
LUNA/3AC (May-Jun 2022)					
h=1	0.163	0.213	0.180	0.167	0.159
h=7	0.159	0.156	0.198	0.183	0.164
FTX (Nov 2022)					
h=1	0.378	0.651	0.416	0.338	0.364
h=7	0.264	0.416	0.270	0.202	0.254
Carry unwind (Aug 2024)					
h=1	0.219	0.340	0.226	0.212	0.217
h=7	0.189	0.301	0.211	0.170	0.185
Liquidation cascade (Oct 2025)					
h=1	0.241	0.236	0.207	0.233	0.239
h=7	0.168	0.131	0.125	0.160	0.167

Forecast origins inside each window; best value per row in bold.

6.4. Training Dynamics

Fig. 4 explains why the selected anchored models remain close to HAR. Without the anchor, the deep models start from validation errors 1.8 to 3.5 times that of HAR and need 25 to 40 epochs to reach their best level, which still lies 4.5% to 37.6% above HAR. With the anchor, the networks start at the HAR level, improve slightly in the first epochs (SOC-Mamba-EF reaches 0.975 at epoch 10) and then over-fit: after 60 epochs the validation error of the anchored models is 25.3% to 49.6% above HAR. With 400 to 1,900 training days, the data support only a small departure from the econometric solution.

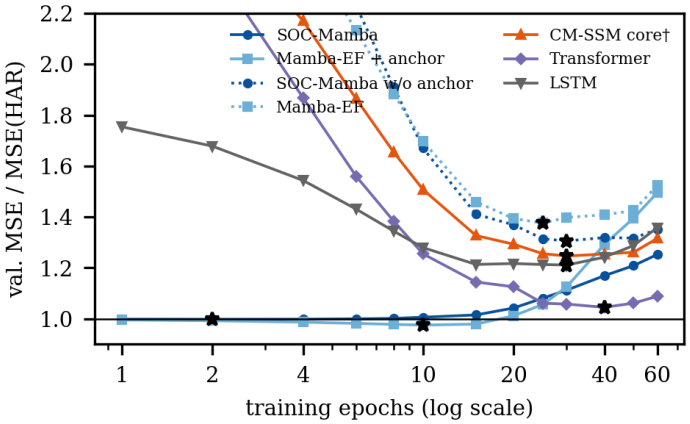


Figure 4. Pooled validation MSE relative to HAR versus training epochs (criterion (12)); stars mark the selected checkpoints. Solid and dotted blue lines show the same networks with and without the HAR anchor.

6.5. Ablation Study

Table V removes one component at a time from SOC-Mamba-XS. Removing the HAR anchor raises the MSE by 18.6%, 32.9% and 34.6% ($p < 0.01$). All other changes, namely removing cross-modal selection, the reliability gate, the asymmetric update, modality dropout or either auxiliary stream, alter the MSE by less than 0.0001. Because the selected checkpoint lies only eight gradient steps away from the zero-initialized cross-modal maps, the auxiliary pathways hardly leave their initial state: the gate weights averaged over the test period are 0.32 (sentiment), 0.37 (on-chain) and 0.31 (null). Under validation-based selection, the XS variant

ISSN: 3080-9843 (Print), ISSN: 3080-9851 (Online)

therefore acts as a lightly corrected HAR model. Separating the contributions of these components at longer training budgets, where all variants over-fit (Fig. 4), calls for stronger regularization.

Table V. Ablation study (pooled test period)

	MSE			QLIKE		
Variant	h=1	h=7	h=30	h=1	h=7	h=30
SOC-Mamba-XS (full)	0.1830	0.1003	0.1129	0.4521	0.2318	0.1813
w/o cross-modal selection	0.1830	0.1003	0.1129	0.4521	0.2318	0.1813
w/o reliability gate	0.1829**	0.1003	0.1129	0.4521	0.2318	0.1812
w/o asymmetric update	0.1830	0.1004*	0.1129	0.4523	0.2317	0.1813
w/o modality dropout	0.1830	0.1003	0.1129	0.4521	0.2318	0.1813
w/o HAR anchor	0.2170***	0.1333***	0.1519***	0.5675	0.3080	0.2496
w/o sentiment stream	0.1829*	0.1003	0.1129	0.4520	0.2317	0.1812
w/o on-chain stream	0.1830	0.1003	0.1129	0.4521	0.2318	0.1812
market stream only	0.1830	0.1004	0.1129	0.4520	0.2317	0.1812

***, **, *: DM test (MSE) of the variant against the full model, $p < 0.01, 0.05, 0.10$.

6.6. Missing Modalities and Prediction Intervals

Table VI simulates feed outages at test time. SOC-Mamba-EF loses accuracy when the auxiliary streams are zero-filled: its average MSE rises by 1.5% without on-chain data and by 2.3% without both streams, whereas removing sentiment changes it by 0.4%. The early-fusion network therefore does use the on-chain information. SOC-Mamba-XS is unaffected, which reflects both its availability mask and its weak reliance on the auxiliary streams at the selected checkpoint.

Table VI. Forecasting with missing modalities at test time

Model	All streams	No sentiment	No on-chain	Neither
SOC-Mamba-XS	0.1321	+0.0%	+0.0%	+0.0%
XS w/o mod. dropout	0.1321	+0.0%	+0.0%	+0.0%
SOC-Mamba-EF	0.1313	+0.4%	+1.5%	+2.3%

First column: MSE averaged over the three horizons with all streams; other columns: relative change of that MSE. Missing streams are zero-filled and, for SOC-Mamba, flagged as unavailable to the gate.

Table VII reports interval calibration. All models under-cover at $h = 1$ and $h = 30$ and over-cover at $h = 7$. The heteroscedastic heads do not improve on HAR with a constant residual variance: SOC-Mamba-EF comes closest to the nominal level at $h = 30$ (85.9% against 84.8%) with slightly wider intervals, and CM-SSM core† produces the widest intervals without better coverage at $h = 30$. Fig. 5 shows 7-day forecasts of SOC-Mamba-EF in 2022 and 2025–2026: the anchored forecast tracks the realized level closely in calm phases but, like HAR, reacts with a lag to abrupt jumps.

ISSN: 3080-9843 (Print), ISSN: 3080-9851 (Online)

Table VII. Central 90% prediction intervals

	Coverage (%)			Mean width		
Model	h=1	h=7	h=30	h=1	h=7	h=30
HAR (homosc.)	87.0	92.3	84.8	1.267	1.073	0.971
CM-SSM core†	88.2	92.1	82.5	1.433	1.236	1.064
SOC-Mamba-EF	87.5	94.1	85.9	1.282	1.119	1.000
SOC-Mamba-XS	87.0	92.4	84.6	1.257	1.087	0.981

Gaussian intervals in log-volatility space; nominal coverage 90%.

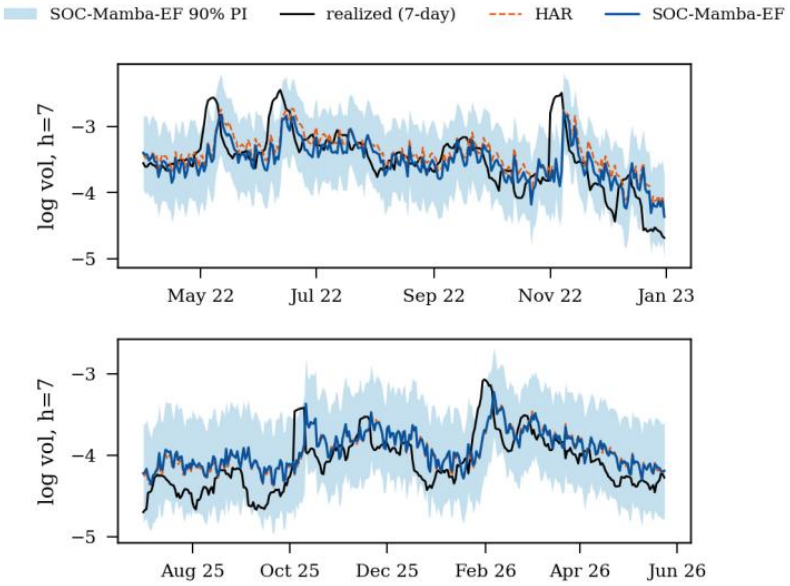


Figure 5. Seven-day volatility forecasts of SOC-Mamba-EF with 90% prediction intervals, HAR forecasts and realized values; top: April–December 2022, bottom: July 2025–May 2026.

6.7. Alternative Volatility Proxy

When the same forecasts are evaluated against realized variance from squared 4-hour returns, the ranking of the leading models changes. The lowest MSE is obtained by HAR-X at h = 1, HAR-X at h = 7 and the random walk at h = 30. SOC-Mamba-EF still beats HAR at every horizon (0.4427, 0.1404 and 0.1269 against 0.4447, 0.1435 and 0.1335), but among the unanchored networks the Transformer does so as well, while the others remain behind HAR. The ranking among the leading models is therefore sensitive to the volatility proxy, whereas the gap between the anchored models and most unanchored networks is not.

7. Discussion

Three lessons follow. First, with one asset and a few years of daily data, the decisive design choice is not the sequence architecture but the relation to the econometric benchmark. Every unanchored network, including a selective SSM with domain-specific selection, was beaten by a four-parameter regression, whereas the same networks with a HAR anchor were at least as accurate as HAR. The advantage of anchoring is clearest on the range-based target; under the squared-return proxy HAR-X ranks first at the 1-day horizon and the Transformer also beats HAR (Section VI-G), so the ordering among the strongest models should be read against the measurement noise of the proxy. Gains reported against weak baselines can overstate the value of new architectures, and comparisons with HAR under a validation-only protocol should be routine.

ISSN: 3080-9843 (Print), ISSN: 3080-9851 (Online)

Second, the evidence on the auxiliary data has to be read separately for the two fusion variants. The early-fusion variant shows episodic benefits from auxiliary information, whereas the cross-modal selective-scan variant provides no clear incremental improvement under the selected evaluation protocol. For SOC-Mamba-EF, the benefit appears in stress periods such as the FTX collapse and in its loss of accuracy when the on-chain stream is removed at test time (Tables IV and VI). For SOC-Mamba-XS, removing either auxiliary stream, the reliability gate or the cross-modal selection changes the MSE by less than 0.0001 (Table V), and its significant 1-day QLIKE gain over HAR (Section VI-B) is matched by its market-only ablation (0.4520 against 0.4521), so this gain stems from the anchored selective core. For neither variant is the full-sample MSE improvement over HAR statistically significant. The predictive relation of these variables is also unstable over time: HAR-X gained at the 30-day horizon in 2022 and 2023 and lost heavily in 2025, when volatility stayed low although MVRV remained elevated.

Third, the cross-modal selective scan, which ties auxiliary information to the memory horizon of the SSM, stayed close to HAR under the selected training budget. Its zero-initialized cross-modal maps and a training budget limited by over-fitting explain this behavior. We expect the mechanism to be more useful when several assets are trained jointly, so that memory modulation can be learned from many more regime changes. Extending the study to further assets, sentiment indices and intraday price feeds, and to additional stress periods, is a natural next step.

8. Conclusion

We proposed SOC-Mamba, a HAR-anchored multimodal selective state space framework for Bitcoin volatility forecasting with sentiment and on-chain data, and evaluated it under a protocol in which every modeling choice is made on validation data. Anchoring turned deep sequence models from clearly worse than HAR into members of the model confidence set, and the early-fusion variant gave the lowest 1-day MSE and the largest gains during stress episodes. Sentiment and on-chain information improved the early-fusion forecasts most during stress episodes, and the selective-scan variant improved significantly on HAR in 1-day QLIKE. Future work will train the framework jointly on several cryptocurrencies, add implied-volatility and order-book information, and test whether memory modulation by exogenous signals becomes beneficial with larger samples.

References

[1] Corbet, S., Lucey, B., Urquhart, A., & Yarovaya, L. (2019). Cryptocurrencies as a financial asset: A systematic analysis. International Review of Financial Analysis, 62, 182–199. https://doi.org/10.1016/j.irfa.2018.09.003

[2] Katsiampa, P. (2017). Volatility estimation for Bitcoin: A comparison of GARCH models. Economics Letters, 158, 3–6. https://doi.org/10.1016/j.econlet.2017.06.023

[3] Corsi, F. (2009). A simple approximate long-memory model of realized volatility. Journal of Financial Econometrics, 7(2), 174–196. https://doi.org/10.1093/jjfinec/nbp001

[4] Bergsli, L. Ø., Lind, A. F., Molnár, P., & Polasik, M. (2022). Forecasting volatility of Bitcoin. Research in International Business and Finance, 59, 101540. https://doi.org/10.1016/j.ribaf.2021.101540

[5] Chen, J., Zhang, X., & Han, X. (2023). Explainable identification of tax-related financial reporting anomalies in U.S. public companies: Cross-fitted tax-consistency residuals and Owen-value Shapley propagation. World Journal of Economics and Business Research, 1(1), 33–45.

[6] Shen, D., Urquhart, A., & Wang, P. (2019). Does Twitter predict Bitcoin? Economics Letters, 174, 118–122. https://doi.org/10.1016/j.econlet.2018.11.007

ISSN: 3080-9843 (Print), ISSN: 3080-9851 (Online)

[7] Jang, H., & Lee, J. (2018). An empirical study on modeling and prediction of Bitcoin prices with Bayesian neural networks based on blockchain information. IEEE Access, 6, 5427–5437. https://doi.org/10.1109/ACCESS.2017.2779181

[8] Gu, A., & Dao, T. (2024). Mamba: Linear-time sequence modeling with selective state spaces. In Proceedings of the Conference on Language Modeling (COLM). https://doi.org/10.48550/arXiv.2312.00752

[9] Gu, A., Goel, K., & Ré, C. (2022). Efficiently modeling long sequences with structured state spaces. In Proceedings of the International Conference on Learning Representations (ICLR). https://doi.org/10.48550/arXiv.2111.00396

[10] Sepehri, M. S., Mehradfar, A., Soltanolkotabi, M., & Avestimehr, S. (2025). CryptoMamba: Leveraging state space models for accurate Bitcoin price prediction. In Proceedings of the IEEE International Conference on Blockchain and Cryptocurrency (ICBC). https://doi.org/10.1109/ICBC64466.2025.11114565

[11] Wang, Z., Kong, F., Feng, S., Wang, M., Yang, X., Zhao, H., Wang, D., & Zhang, Y. (2025). Is Mamba effective for time series forecasting? Neurocomputing, 619, 129178. https://doi.org/10.1016/j.neucom.2024.129178

[12] Bollerslev, T. (1986). Generalized autoregressive conditional heteroskedasticity. Journal of Econometrics, 31(3), 307–327. https://doi.org/10.1016/0304-4076(86)90063-1

[13] Glosten, L. R., Jagannathan, R., & Runkle, D. E. (1993). On the relation between the expected value and the volatility of the nominal excess return on stocks. The Journal of Finance, 48(5), 1779–1801. https://doi.org/10.1111/j.1540-6261.1993.tb05128.x

[14] Garman, M. B., & Klass, M. J. (1980). On the estimation of security price volatilities from historical data. Journal of Business, 53(1), 67–78. https://doi.org/10.1086/296072

[15] Martens, M., & van Dijk, D. (2007). Measuring volatility with the realized range. Journal of Econometrics, 138(1), 181–207. https://doi.org/10.1016/j.jeconom.2006.05.019

[16] Zhou, F. (2026). Moment-constrained side-force allocation for wind-assisted merchant ships: Hydrodynamic performance of retractable anti-leeway fins and flap rudders. Innovation and Technology Studies, 3(1), 27–41.

[17] Zhao, X., Liu, J., Wang, Y., & Wang, J. (2026). CryptoMamba-SSM: Linear complexity state space models for cryptocurrency volatility prediction. IEEE Open Journal of the Computer Society, 7, 226–243. https://doi.org/10.1109/OJCS.2026.3651226

[18] Baltrušaitis, T., Ahuja, C., & Morency, L.-P. (2019). Multimodal machine learning: A survey and taxonomy. IEEE Transactions on Pattern Analysis and Machine Intelligence, 41(2), 423–443. https://doi.org/10.1109/TPAMI.2018.2798607

[19] Neverova, N., Wolf, C., Taylor, G., & Nebout, F. (2016). ModDrop: Adaptive multi-modal gesture recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence, 38(8), 1692–1706. https://doi.org/10.1109/TPAMI.2015.2461544

[20] Alternative.me. (2026). Crypto Fear & Greed Index. https://alternative.me/crypto/fear-and-greed-index/

[21] Coin Metrics. (2026). Coin Metrics community network data [Data set]. https://github.com/coinmetrics/data

[22] Yue, X., & Zhu, R. (2024). An explainable hybrid machine learning framework for cash-flow-conditioned multi-horizon liquidity risk early warning: An SME-oriented benchmark study. World Journal of Management Science, 2(1), 63–72.

[23] Loshchilov, I., & Hutter, F. (2019). Decoupled weight decay regularization. In Proceedings of the International Conference on Learning Representations (ICLR). https://doi.org/10.48550/arXiv.1711.05101

[24] Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., & Liu, T.-Y. (2017). LightGBM: A highly efficient gradient boosting decision tree. In Advances in Neural Information Processing Systems (Vol. 30, pp. 3146–3154).

ISSN: 3080-9843 (Print), ISSN: 3080-9851 (Online)

[25] Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. Neural Computation, 9(8), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

[26] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. In Advances in Neural Information Processing Systems (Vol. 30, pp. 5998–6008).

[27] Patton, A. J. (2011). Volatility forecast comparison using imperfect volatility proxies. Journal of Econometrics, 160(1), 246–256. https://doi.org/10.1016/j.jeconom.2010.03.034

[28] Diebold, F. X., & Mariano, R. S. (1995). Comparing predictive accuracy. Journal of Business & Economic Statistics, 13(3), 253–263. https://doi.org/10.1080/07350015.1995.10524599

[29] Harvey, D., Leybourne, S., & Newbold, P. (1997). Testing the equality of prediction mean squared errors. International Journal of Forecasting, 13(2), 281–291. https://doi.org/10.1016/S0169-2070(96)00719-4

[30] Hansen, P. R., Lunde, A., & Nason, J. M. (2011). The model confidence set. Econometrica, 79(2), 453–497. https://doi.org/10.3982/ECTA5771

[31] Zhang, X., Yang, Y., & Wang, Y. (2026). Risk-Aware Decision-Focused Demand Forecasting for Multi-Channel Inventory Optimization under Promotion and Stock-Out Constraints. Highlights in Business, Economics and Management, 69, 263–279.

[32] Ge, Y., Wang, Y., Liu, J., & Wang, J. (2025). GAN-Enhanced Implied Volatility Surface Reconstruction for Option Pricing Error Mitigation. IEEE Access. https://doi.org/10.1109/ACCESS.2025.3619553