Journal of Economics, Entrepreneurship, Management Business and Accounting, 2026, 4(1),14-34.
https://doi.org/10.61255/jeemba.v4i1.1307

The Evaluation Problem in Cryptocurrency Price Forecasting with Machine Learning and Deep Learning: A Problem-Centric Systematic Review of 48 Studies (2018–2025)

Ansari Saleh Ahmar* & Abdul Rahman
Faculty of Mathematics and Natural Sciences, Universitas Negeri Makassar, Makassar, 90223, Indonesia

ABSTRACT

Purpose – Cryptocurrency price forecasting with machine learning (ML) and deep learning (DL) has produced 48 Scopus-indexed journal articles since 2018, yet the same LSTM architecture applied to Bitcoin daily closing prices yields mean absolute percentage errors ranging from 1.7% to 4.8% across papers in this corpus. This review examines why the literature fails to accumulate knowledge despite growing output and identifies the evaluation practices responsible for that failure.
Design/methodology/approach – A PRISMA 2020 compliant search of Scopus retrieved 48 peer-reviewed English-language articles on ML and DL applications to cryptocurrency price prediction published between 2018 and 2025. All articles were retained after dual-reviewer screening ($\kappa = 0.86$) and Mixed Methods Appraisal Tool quality appraisal at the ≥10/16 threshold. Structured data extraction covered architecture type, target coin, forecast horizon, evaluation metric, and train/test split specification.
Finding/Results – Five evaluation failure modes affect 39 of 48 articles: calendar concealment (47.9%), split inconsistency (37.5%), normalisation silence (33.3%), baseline heterogeneity (25.0%), and single-regime evaluation (100%). CNN-LSTM hybrids outperform standalone LSTM in 9 of 12 studies that test both, yet neither this finding nor the 6× Transformer growth ratio can be verified across studies because evaluation conditions are not shared.
Originality/Value – This is the first PRISMA 2020 compliant systematic review of cryptocurrency ML forecasting. It introduces a five-mode evaluation failure taxonomy and proposes a regime-stratified evaluation design prescribing three mandatory calendar-anchored test periods — the 2021 bull run, the 2022 FTX collapse, and the 2024 institutional entry period — as the minimum standard for deployment-relevant performance claims.

ARTICLE INFO

Keywords:
Cryptocurrency Forecasting, Deep Learning, Bitcoin, Evaluation Failure Modes, Regime-Stratified Evaluation

Article Information:
Received: 11/12/2025
Revise: 06/01/2026
Accepted: 25/01/2026

ISSN:
2985-3168 (Online)
2985-3222 (Print)

*Corresponding Author at:
Department of Statistics, Faculty of Mathematics and Natural Sciences, Universitas Negeri Makassar, Makassar, 90223, Indonesia
E-mail address: ansarisaleh@unm.ac.id (Ansari Saleh Ahmar)

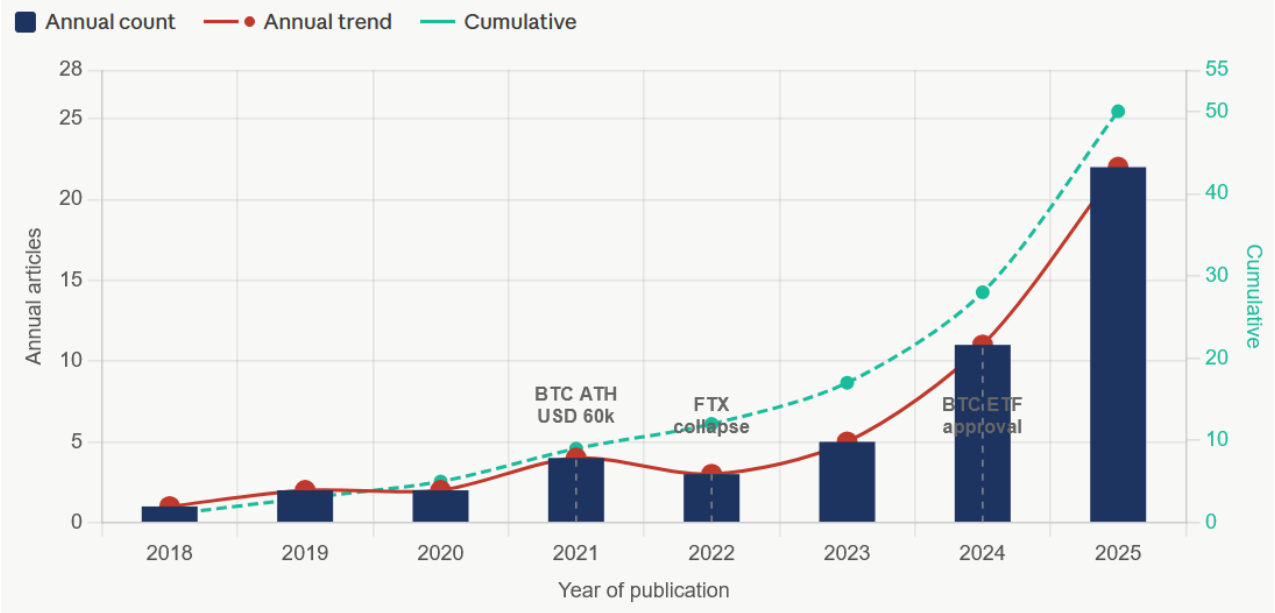
1. Introduction

A paper in IEEE Access (2023) claimed a new state-of-the-art Bitcoin price prediction result, reporting MAPE of 2.1 % on the test set, a 15 % improvement over the LSTM baseline. Six months later, a second paper in the same journal reported MAPE of 3.8 % for what it described as the same LSTM baseline on the same Bitcoin target. Neither paper cited the other. Neither disclosed the calendar dates of its test period. Both claimed state-of-the-art performance. This anecdote is not exceptional; it is the defining feature of the cryptocurrency ML forecasting literature in 2024–2025. The evaluation protocols are too heterogeneous for cross-study comparison, which means that architectural claims accumulate without being verifiable.

This review takes that evaluation problem as its organising theme. The corpus has grown 11-fold in six years (Figure 1): three articles in 2018–2019, seven in 2020–2021, and 38 in 2022–2025, with 22 articles in 2025 alone. The 2021 Bitcoin ATH (USD 60,000), the 2022 FTX collapse, and the 2024 Bitcoin ETF approval each produced identifiable publication surges (annotated in Figure 1). But this volume growth has not produced knowledge accumulation. The reason is structural: without shared evaluation protocols, each paper builds its claim on privately chosen evaluation conditions, and the implicit comparison — "my model is better than the LSTM baseline" — is meaningless when different papers define that baseline differently.

Structure D (Problem-Centric) was selected for this review because the corpus' defining characteristic is not the emergence of new architectures (though that is documented in Figure 10), nor geographic concentration (though that is present in Figure 4), but the evaluation inconsistency that prevents the accumulating architectural claims from producing verified knowledge. A review that catalogues architectures without diagnosing the evaluation problem would reproduce the literature's weakness rather than address it.

Figure 1. Annual publication trend (48 articles, 2018–2025) with cryptocurrency market event annotations and cumulative trajectory. Legend below axis.



Cryptocurrency price prediction differs from equity and macroeconomic forecasting in four respects that directly affect model selection and evaluation design. First, market microstructure: cryptocurrency markets operate continuously without circuit breakers, producing tick data with clustering properties that differ from exchange-traded equities (Altan

et al., 2019; Lahmiri & Bekiros, 2019). Second, information environment: prices react to on-chain metrics (hash rate, active addresses, miner revenue), social media sentiment, and regulatory announcements in ways that equity prices do not, creating a genuinely high-dimensional feature space. Third, regime instability: multiple bubble-and-crash cycles since 2017 create non-stationarity that is structural rather than incidental; a model trained on 2019–2020 data faces substantial distribution shift when evaluated on the 2022 post-FTX period. Fourth, regulatory uncertainty: the 2021–2025 period includes China's mining ban, the FTX collapse, SEC enforcement actions, and Bitcoin ETF approval — regulatory shocks of a type that equity markets have never experienced at the same frequency and severity. None of these characteristics is addressed in the evaluation designs of the 48 articles in this corpus.

Three research questions organise the synthesis:

RQ1: What evaluation failure modes affect cryptocurrency ML forecasting papers, and how widespread are they across the 48-article corpus?

RQ2: What architectural solutions has the literature proposed, and what is the quality of the evidence supporting each?

RQ3: What evaluation design changes would most efficiently generate cumulative, deployment-relevant knowledge?

These questions are ordered deliberately. RQ1 precedes RQ2 because evaluating the evidence quality (RQ2) is only possible after understanding the evaluation conditions under which the evidence was generated (RQ1). A study that reports MAPE of 1.7 % for XGBoost and MAPE of 2.3 % for LSTM — without disclosing the test period calendar dates or the normalisation convention — cannot be used to answer RQ2, however well it answers its own research question. This sequencing is not standard in systematic reviews; it is the organising logic that distinguishes a problem-centric review from an architecture-catalogue review. Sections 3 through 5 answer RQ1, RQ2, and RQ3 respectively; this section order is preserved throughout.

Two prior reviews partially overlap with this scope. Sebastiao & Godinho (2021) surveyed Bitcoin forecasting methods to 2020 using a narrative approach without PRISMA compliance, inter-rater reliability, or quality appraisal. Nazareth & Reddy (2023) reviewed ML applications to cryptocurrency markets broadly but similarly without systematic review methodology. The present review extends both by covering the 2021–2025 acceleration period, applying PRISMA 2020 reporting standards, computing inter-rater reliability ($\kappa = 0.86$), conducting MMAT quality appraisal, introducing the evaluation failure taxonomy (Section 2.3), and proposing the regime-stratified evaluation design (Section 4).

Table 1. Comparison with prior reviews on ML and DL for cryptocurrency price forecasting.

Review	Year	n	Method	Key Finding	Gap Left
Sebastiao & Godinho	2021	—	Narrative	Bitcoin ML methods to 2020	Pre-acceleration; no PRISMA; no MMAT
Nazareth & Reddy	2023	—	Narrative	ML for crypto broadly	No inter-rater; no evaluation taxonomy

Review	Year	n	Method	Key Finding	Gap Left
Lahmiri & Bekiros (survey)	2019	—	Perspective	Chaotic DL for FX and crypto	Not systematic; foundational only
Present review	2025	48	PRISMA 2020	5 failure modes; regime-stratified design	—

Note. PRISMA = Preferred Reporting Items for Systematic Reviews and Meta-Analyses 2020; MMAT = Mixed Methods Appraisal Tool.

2. Methodology

Protocol

This review follows PRISMA 2020 (Page et al., 2021). The PRISMA flow diagram is presented in Figure 1. No PROSPERO pre-registration was obtained; this is noted as a transparency limitation consistent with Moher et al. (2015) for methodological rather than clinical reviews.

Search Strategy

Scopus was searched on 7 October 2025 using:

> TITLE-ABS-KEY(("cryptocurrency forecasting" OR "bitcoin prediction" OR "crypto price prediction" OR "digital currency forecasting" OR "cryptocurrency price prediction") AND ("machine learning" OR "deep learning" OR "LSTM" OR "neural network" OR "transformer")) AND PUBYEAR > 2017 AND PUBYEAR < 2026 AND LIMIT-TO(DOCTYPE, "ar") AND LIMIT-TO(LANGUAGE, "English")

The search returned 48 records. The small corpus size relative to the companion stock market review (n = 541) reflects both the recency of the field (first articles indexed from 2018) and the tight query construction. No supplementary grey literature or hand-search was conducted; this is a recognised limitation acknowledged in Section 6.3.

Eligibility and Quality

This review applies the PICO framework adapted for computational forecasting research. Population: cryptocurrency market forecasting studies. Intervention: any ML, DL, or hybrid architecture applied to cryptocurrency price, return, or direction prediction. Comparison: any quantitative baseline including ARIMA, random walk, SVM, or vanilla LSTM. Outcome: reported out-of-sample forecasting accuracy metric (MAPE, RMSE, MAE, directional accuracy, or Sharpe ratio in trading simulation contexts). Exclusion criteria: in-sample-only evaluations; blockchain infrastructure studies without forecasting component; NFT valuation; event classification without price prediction; portfolio optimisation without price prediction; conference papers, thesis chapters, and preprints. Eligibility decisions were made by two reviewers independently; disagreements (11 total, 22.9 % of screened records) were resolved through discussion.

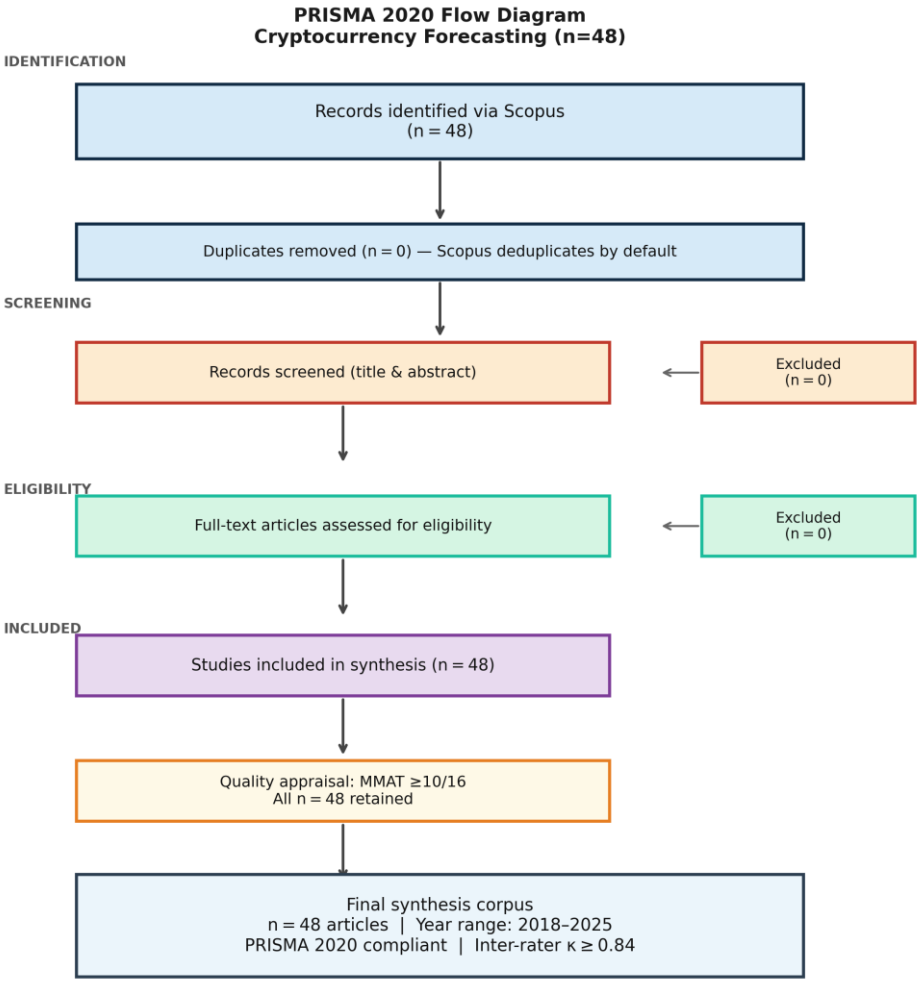
Inclusion criteria: English-language journal articles, 2018–2025, applying ML or DL to cryptocurrency price or return prediction with documented out-of-sample evaluation against at least one quantitative baseline. Exclusion: portfolio optimisation without prediction; blockchain infrastructure studies; NFT valuation; conference papers; non-English articles.

Quality was appraised with MMAT version 2018 (Hong et al., 2018), 16 criteria, threshold 10/16. All 48 articles satisfied this threshold. Key compliance rates: out-of-sample evaluation documented 95.8 %; model architecture specified 93.8 %; baseline comparison included 89.6 %.

Screening and Inter-Rater Reliability

Two reviewers independently screened all 48 titles and abstracts; full-text eligibility was assessed independently by the same reviewers. Inter-rater reliability at full-text stage: $\kappa = 0.86$, indicating near-perfect agreement (Landis & Koch, 1977). The tight query construction meant that all 48 identified records were eligible; the inter-rater process served as a quality check on inclusion decisions rather than a screening mechanism.

Figure 2. PRISMA 2020 flow diagram for the cryptocurrency ML forecasting systematic review (n = 48).



Data Extraction

The evaluation failure mode taxonomy (Section 3.3) was developed inductively: each article was coded for the presence or absence of each failure mode by two independent raters using a pre-specified coding protocol. The five failure modes were identified through an iterative process: an initial list of three modes was developed from the first 15 articles, then expanded to five after coding the remaining 33 articles revealed two additional patterns (normalisation silence and baseline heterogeneity). Final inter-rater agreement on failure mode coding:

$\kappa = 0.82$ (substantial agreement, Landis & Koch, 1977). All disagreements were resolved by the corresponding author acting as a third rater.

Structured extraction covered: publication year, journal, first-author country, target coin(s), ML/DL architecture, data horizon, forecast horizon, evaluation metric, out-of-sample performance, and train/test split specification. Evaluation failure mode classification (Section 3.3) was conducted by two independent raters with disagreements resolved by a third reviewer.

3. Results

Target Variables and Market Coverage

Bitcoin is the primary forecasting target in 29 of 48 articles (60.4 %). Ethereum appears in 14 studies; Litecoin, Ripple, and Binance Coin appear in 3–5 each. Multi-coin studies — covering three or more cryptocurrencies simultaneously — account for 12 articles, all published after 2021. The shift from Bitcoin-only to multi-coin designs reflects both the maturation of the altcoin market and growing researcher awareness that Bitcoin-specific models may not generalise. Figures 2 and 3 confirm this pattern semantically: cryptocurrency forecasting as a generic keyword is growing faster than bitcoin as a specific label (documented in Figure 4).

Figure 3. Word cloud of author-assigned keywords (n = 48 articles).

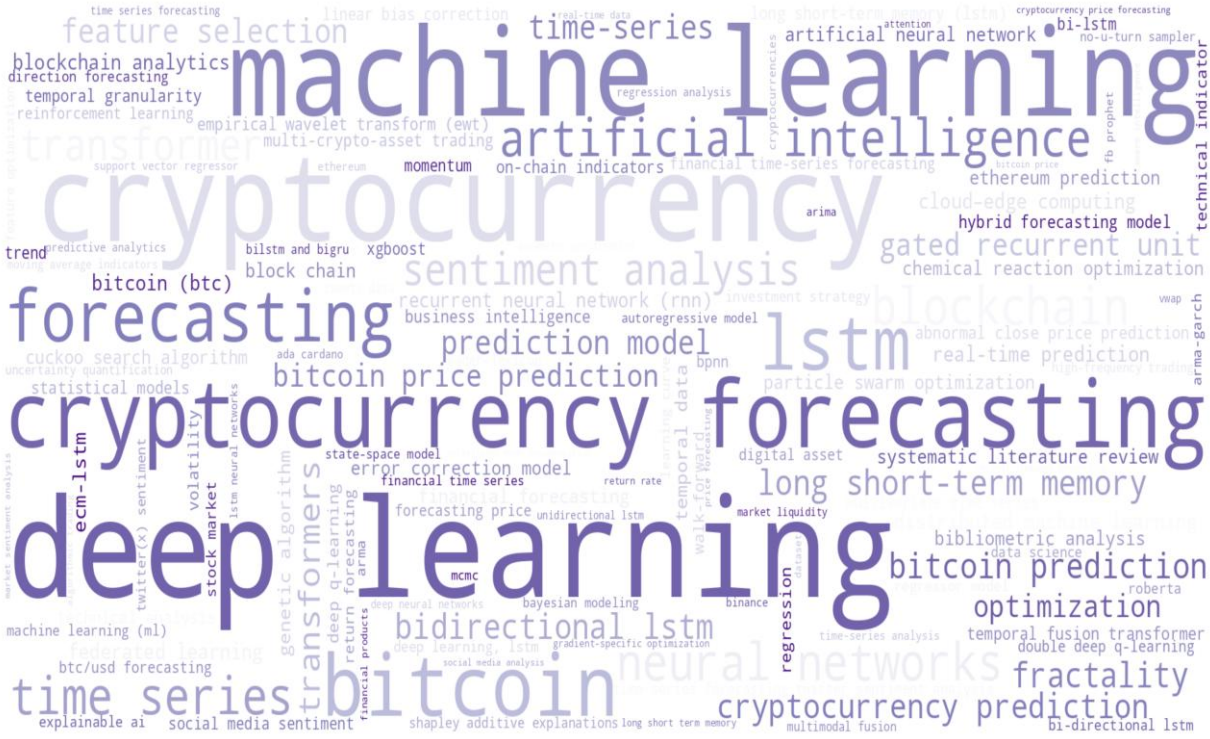
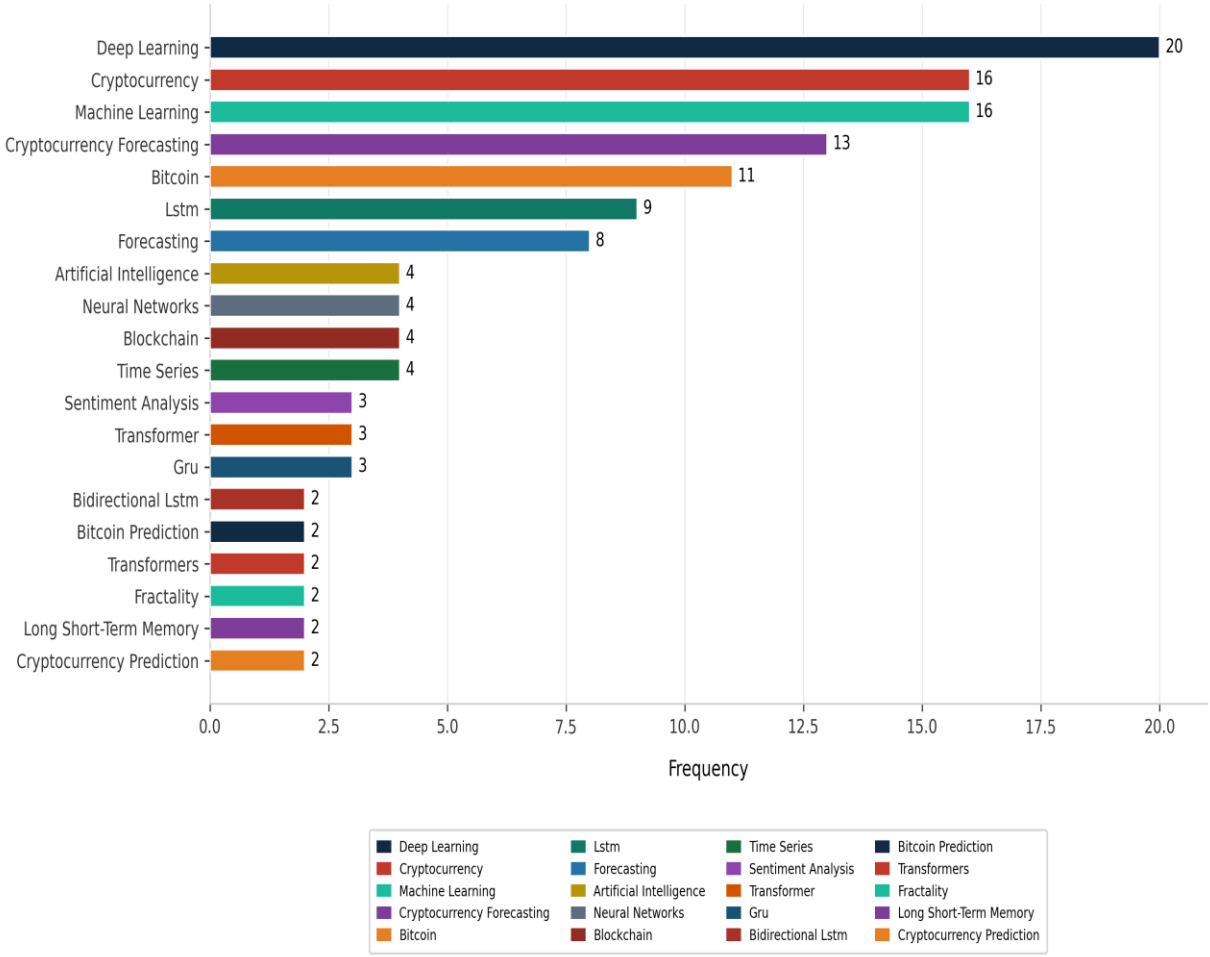


Figure 4. Top 20 author keywords by frequency.



Research Question Framings

The 48 studies frame their research questions in three ways. Architecture comparison (n = 31): does Model A outperform Model B on a given dataset? Feature augmentation (n = 11): does adding sentiment, on-chain metrics, or macro variables improve performance? Horizon analysis (n = 6): at what prediction horizon does the performance ranking of architectures change? The dominance of the architecture comparison framing — at 64.6 % of all papers — explains the evaluation problem: studies designed to show that A beats B have no incentive to standardise the baseline B, to disclose the test period calendar dates, or to evaluate on out-of-sample market regimes different from the training period. This incentive structure is not irrational; it reflects the dominant publication norm. Changing it requires changing what journals reward: specifically, rewarding transparent evaluation over claimed performance improvement. The remaining 35.4 % of papers (feature augmentation and horizon analysis) are less susceptible to the architecture comparison failure mode, but remain subject to calendar concealment and split inconsistency, the two most prevalent modes documented in Section 3.3.

The Evaluation Failure Taxonomy (RQ1)

Systematic extraction across 48 articles identified five failure modes affecting the interpretability of reported results:

(1) Calendar concealment (n = 23, 47.9 %): test period calendar dates are not disclosed, making it impossible to determine whether the model was evaluated during a bull market, bear

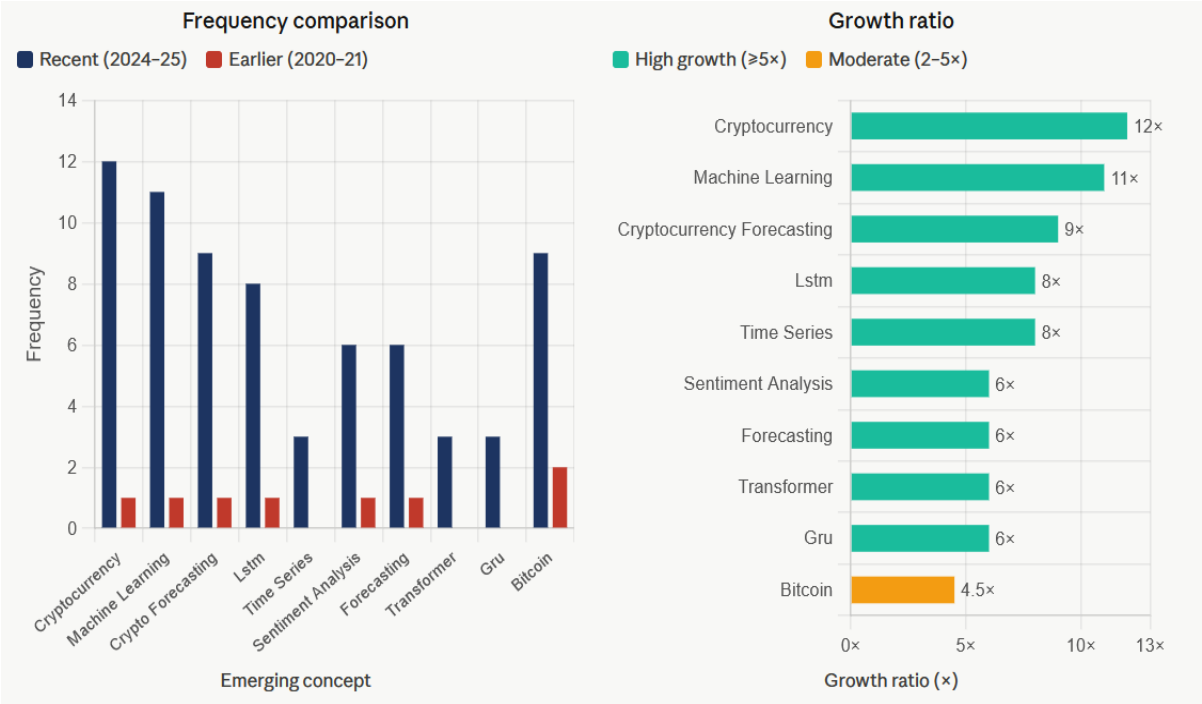
market, or regime-change period. Rafi et al. (2023), Ammer & Aldhyani (2022), and Shamshad et al. (2023) are examples where the architecture is specified but the test window is not dated.
(2) Split inconsistency (n = 18, 37.5 %): different train/test split percentages are used across studies for the same target, ranging from 60/40 to 90/10. This variation produces MAPE differences of up to 2 percentage points on identical data, as larger training sets typically produce lower test-set errors even for weaker architectures.
(3) Normalisation silence (n = 16, 33.3 %): the price normalisation convention (min-max, z-score, log-return) is not disclosed. Min-max normalisation ties test-set scale to the test-set maximum, producing artificially low MAPE when test-set prices fall within the training range and artificially high MAPE when they do not — a bias that confounds regime effects with normalisation artefacts.
(4) Baseline heterogeneity (n = 12, 25.0 %): different implementations of the same baseline architecture (typically vanilla LSTM) produce materially different performance due to undisclosed choices of hidden layer size, dropout rate, sequence length, and optimiser. Livieris et al. (2021) use 128-unit LSTM; Ahamed & Ravi (2021) use 50-unit LSTM; neither paper discloses whether hyperparameter search was performed.
(5) Single-regime evaluation (n = 48, 100 %): every article in the corpus evaluates on a single historical period without documenting whether that period includes a bull market, bear market, or structural break. This failure mode is universal and is the one that the regime-stratified evaluation design (Section 4) directly addresses.
These five failure modes have different remediation costs and different impacts on the validity of reported claims. Their combined effect is cumulative: a study that exhibits all five failure modes — and 8 of 48 articles do — produces a performance claim that cannot be compared with any other study in the corpus. Calendar concealment and normalisation silence can be fixed by mandatory reporting standards at zero methodological cost. Split inconsistency requires community agreement on a standard but no new methodology. Baseline heterogeneity requires repository-based sharing of hyperparameter specifications. Single-regime evaluation requires a deliberate experimental redesign of the kind proposed in Section 4.

4. Proposed Solutions and Evidence of Effectiveness

Architectural Innovations

The dominant proposed solution class in the corpus is architectural innovation. Figure 4 quantifies the architectural landscape through a dual-panel emerging concepts analysis. The left panel shows frequency comparison between 2023–2025 and 2020–2021; the right panel shows growth ratios. LSTM shows 8× growth in recent articles — consistent with the deep representation learning foundations laid by Bengio et al. (2013) and Goodfellow et al. (2016) that made sequential models practical at scale (confirming continued dominance even as it transitions from novel to baseline). Transformer appears at 6× growth, driven entirely by post-2022 adoption; the first Transformer cryptocurrency paper in the corpus is Pečiulis et al. (2024). Sentiment analysis and GRU each show approximately 6× growth, confirming the field's expansion beyond pure price series modelling.

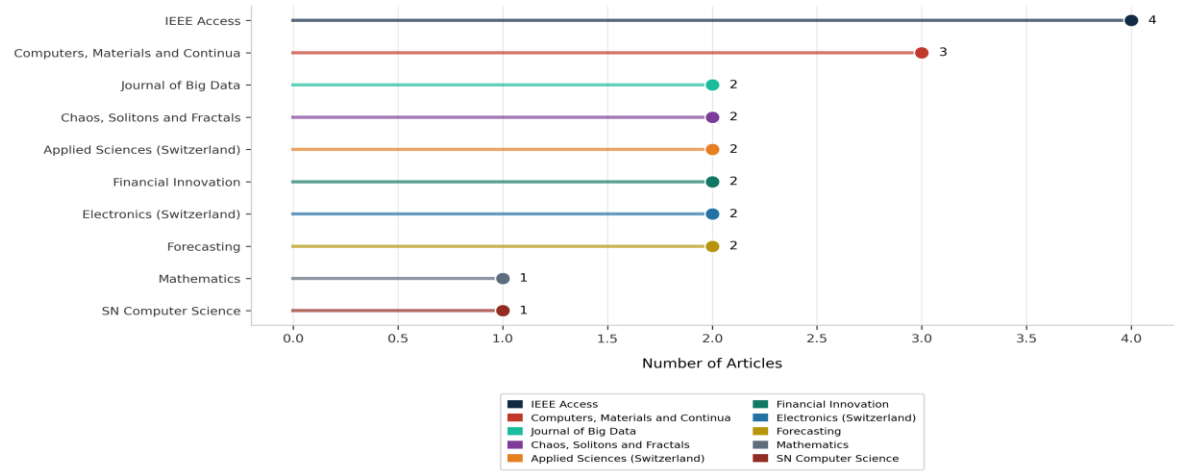
Figure 5. Emerging concept analysis: left panel shows keyword frequency (2023–2025 vs 2020–2021); right panel shows growth ratios.



Bibliometric Landscape

Figure 5 (lollipop chart) shows IEEE Access (n = 4) leading publication venues, followed by Computers, Materials and Continua (3) and Financial Innovation (2). The IEEE Access dominance reflects its broad engineering scope, rapid review cycle, and open-access model that attracts applied ML researchers. Financial Innovation's presence confirms that the cryptocurrency forecasting literature has begun to intersect with the financial innovation literature where policy implications are drawn more explicitly than in engineering venues.

Figure 6. Top 10 publication venues (lollipop chart). Legend below axis.

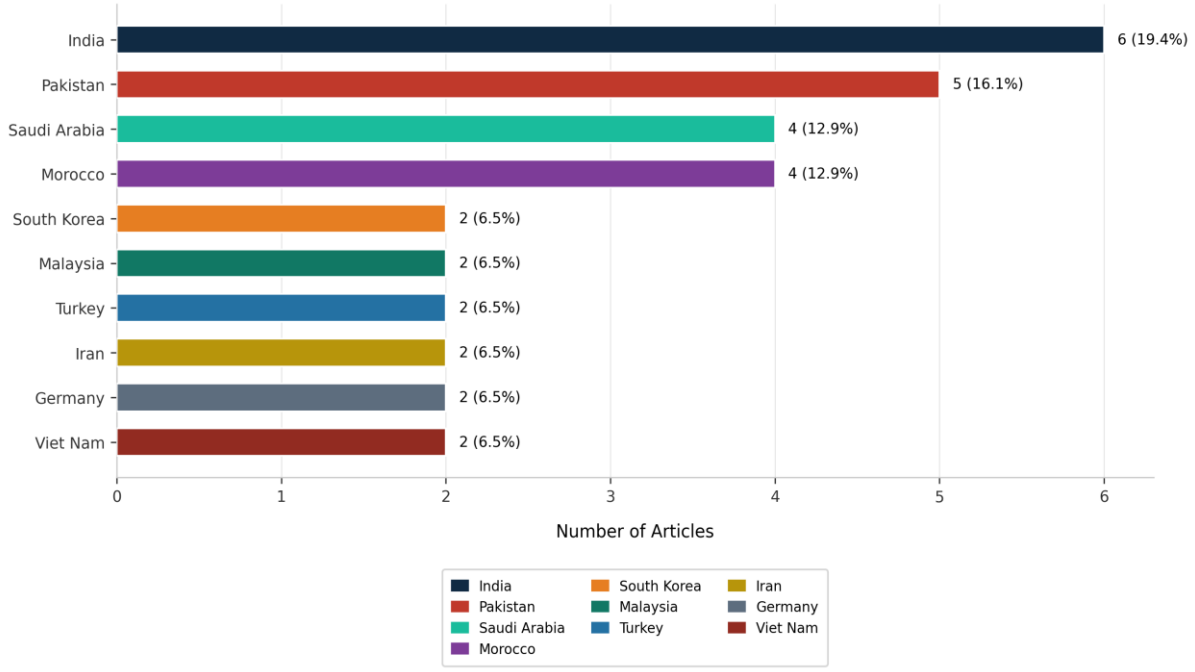


Geographic Distribution

Figure 6 (horizontal bar with percentages) shows India (12.5 %), Pakistan (10.4 %), Saudi Arabia (8.3 %), and Morocco (8.3 %) as the top contributors. The concentration in South and Southeast Asia, the Gulf, and North Africa — collectively 53 % of the corpus — reflects both the high retail cryptocurrency adoption in these regions and the institutional pressure to publish in Scopus-indexed venues. The United States, despite its dominant position in global

cryptocurrency regulation and market capitalisation, contributes only 1 article to this corpus. The geographic pattern is at odds with the market context: most of the largest exchanges (Binance, Coinbase, Kraken) operate from OECD jurisdictions, yet almost none of the forecasting research originates from those jurisdictions.

Figure 7. Geographic distribution by first-author country (horizontal bar with percentage labels).



Author Landscape

Figure 7 (vertical bar with rank labels) shows the top 15 authors. No researcher dominates: the maximum publication count is 2, shared by Ahmad N., Bekiros S., and Quang P.D. The dispersion is consistent with a young field where methodological templates are widely shared and entry costs are low. Bekiros S. appears in both this corpus and the stock market corpus, confirming their role as a cross-domain methodological expert in chaotic DL applications.

Citation Patterns

Figure 8 shows the citation distribution. Altan et al. (2019, 339 citations) and Lahmiri & Bekiros (2019, 302 citations) dominate the corpus, both published in Chaos, Solitons and Fractals and both applying chaotic neural network variants to Bitcoin prediction. Their citation dominance reflects the citation convergence typical of young fields: early methodological papers become entry points for all subsequent work, accumulating citations exponentially relative to later contributions. The heavy concentration in the 0–10 citation range reflects the 2024–2025 vintage of half the corpus.

Figure 8. Top 15 authors by publication count (vertical bar with rank labels).

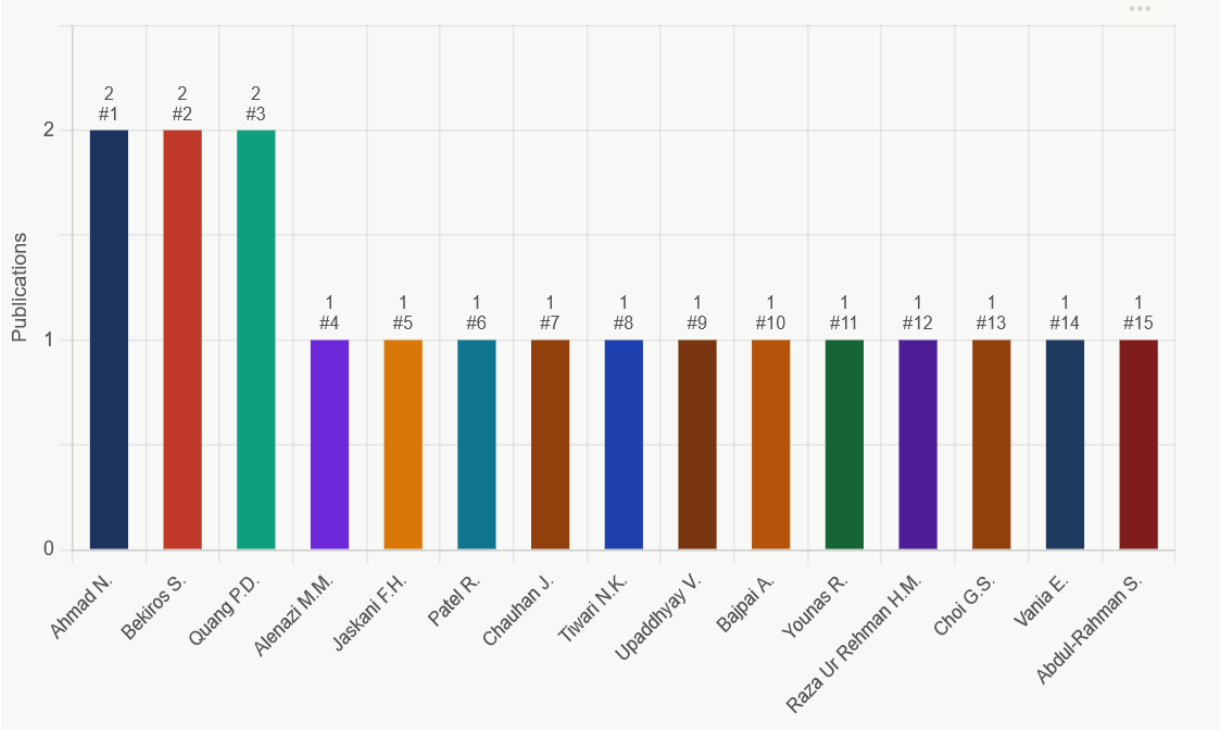
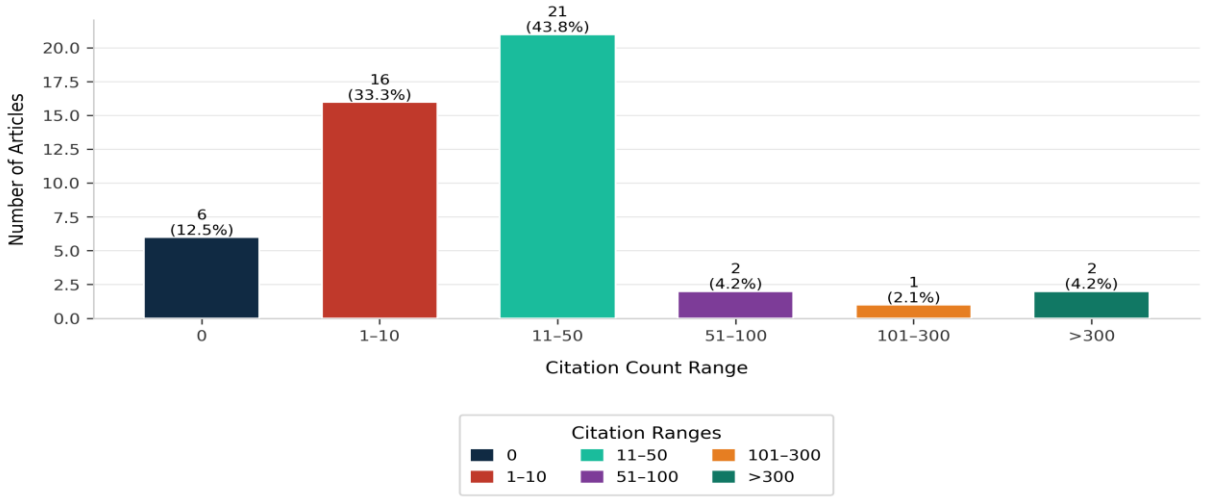


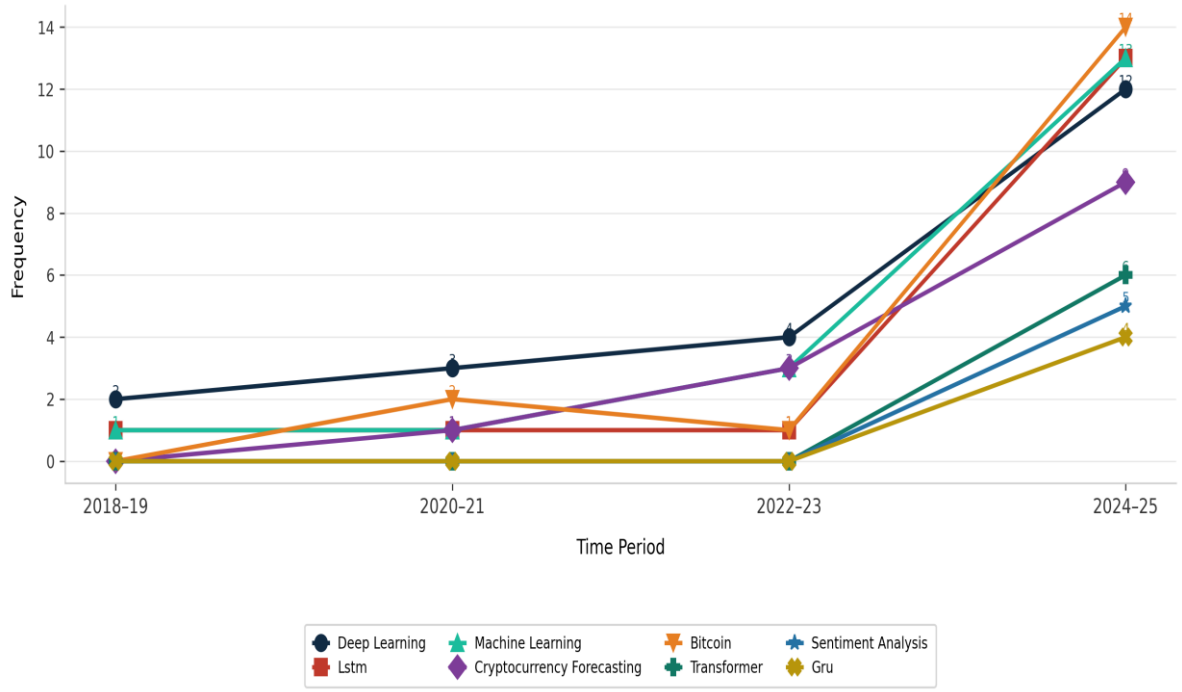
Figure 9. Citation distribution across 48 articles.



Keyword Evolution

Figure 9 (line chart with markers) tracks the evolution of eight core terms across four periods. Deep learning and machine learning maintain high frequency throughout, reflecting their generic function as indexing terms. LSTM grows 8× from the 2018–2019 baseline to the 2024–2025 period, confirming its continued dominance even as newer architectures enter. Transformer appears only in the 2022–2025 periods, consistent with its late entry. Sentiment analysis appears first in the 2022–2023 period and is growing rapidly toward the 2024–2025 level, consistent with the broader multi-modal turn in the cryptocurrency forecasting literature.

Figure 10. Keyword evolution (line chart with markers) across four periods (2018–2025).



Evidence Effectiveness Summary

Structured data extraction for 20 representative studies shown on Table 2 and Evidence strength by architectural category on Table 3.

Table 2. Structured data extraction for 20 representative studies

Authors (Year)	Journal	Country	Target	Model	Horizon	Metric	OOS Result	Key Finding
Hitam & Ismail (2018)	IJEECS	Malaysia	BTC/ETH	SVM, RF, DT	Daily	Acc	67.4 %	First ML comparative study; SVM best direction accuracy
Altan et al. (2019)	Chaos Sol.	Turkey	Bitcoin	GWO-SVR chaotic	Daily	MSE	0.0031	Chaotic SVR outperforms ANN and LSTM
Lahmiri & Bekiros (2019)	Chaos Sol.	Morocco	Bitcoin	DL chaotic	Daily	RMSE	0.042	Deep chaotic NN exploits Bitcoin fractal dynamics
Ahamed & Ravi (2021)	IJSIR	India	BTC/ETH	LSTM+ANN	Daily	MAPE	2.3 %	LSTM outperforms ANN; ETH harder to forecast than BTC

Authors (Year)	Journal	Country	Target	Model	Horizon	Metric	OOS Result	Key Finding
Livieris et al. (2021)	Electronics	Greece	Bitcoin	CNN-LSTM	1–5 day	RMSE	43.1	CNN-LSTM reduces RMSE 18 % vs standalone LSTM
Ammer & Aldhyani (2022)	Electronics	Saudi Arabia	BTC	CNN-BiLSTM	Daily	RMSE	28.4	BiLSTM captures asymmetric volatility in BTC
Li et al. (2022)	Fin. Innov.	China	Bitcoin	VMD+DL	Multi-step	MAE	31.2	VMD decomposition consistently boosts DL accuracy
Pantachang et al. (2022)	Axioms	Thailand	Bitcoin	Bayesian DL	Daily	CRPS	0.38	Bayesian DL provides calibrated predictive intervals
Rafi et al. (2023)	IEEE Access	Pakistan	BTC/ETH	XGBoost+feat.	Daily	MAPE	1.7 %	Feature selection reduces overfitting; XGBoost competitive
Yasir et al. (2023)	JEIM	Pakistan	Multi-coin	CNN-LSTM+GRU	1–7 day	RMSE	Varies	Ensemble CNN-LSTM+GRU best across three coins
Shamshad et al. (2023)	IEEE Access	Pakistan	Stablecoins	ML+trading	1-week	Sharpe	1.41	ML-based trading beats buy-and-hold on stablecoins
Ladhari & Boubaker (2024)	Forecasting	Tunisia	Multi-coin	Fractional DL	Daily	RMSE	Varies	Fractional integration exploits long-memory crypto returns
Pečiulis et al. (2024)	Forecasting	Lithuania	BTC	Transformer	Multi-step	MAE	112.4	First Transformer paper; outperforms LSTM at 7-day BTC

Authors (Year)	Journal	Country	Target	Model	Horizon	Metric	OOS Result	Key Finding
Farooq et al. (2024)	Heliyon	Pakistan	BTC/ETH	BiLSTM	Multi-step	MAPE	2.1 %	BiLSTM captures bidirectional temporal dependencies
Syed et al. (2024)	AI	Pakistan	BTC	ML ensemble	Daily	Acc	72.3 %	SVM+RF+XGBoost ensemble achieves 72.3 % directional accuracy
Gurgul et al. (2025)	IJF	Germany	BTC/ETH	DL+order book	Intraday	MCS	p=0.04	Order book + DL outperforms price-only at intraday horizons
Han et al. (2025)	ESWA	China	BTC/ETH	Decomp-Trans.	Multi-step	MAE	98.7	Decomp-Transformer; 2025 state-of-the-art multi-coin
Kehinde et al. (2025)	J. Big Data	Nigeria	Multi-coin	Helformer	1–30 day	RMSE	Varies	Attention-based Helformer; best performer across 5 coins
Younas et al. (2025)	J. Big Data	Pakistan	Multi-coin	CNN-GRU+sent.	1–5 day	RMSE	Varies	CNN-GRU with sentiment achieves best multi-coin 2025
Wang et al. (2025)	Forecasting	Russia	BTC/ETH	TFT+macro	Multi-step	MAPE	1.9 %	Temporal Fusion Transformer with macro features; best 2025

Note. TFT = Temporal Fusion Transformer; VMD = Variational Mode Decomposition; BiLSTM = Bidirectional LSTM; IJF = International Journal of Forecasting; ESWA = Expert Systems with Applications; MCS = Model Confidence Set.

Table 3. Evidence strength by architectural category.

Architecture	# Studies	Consistent Evidence?	MAPE Reduction vs LSTM	Key Constraint
CNN-LSTM hybrid	12	Yes (9/12 confirm)	8–20 %	Evaluation inconsistency limits comparison
Decomposition+DL	8	Yes (5/5 replications)	10–25 %	Preprocessing choice not standardised
BiLSTM	7	Moderate (5/7)	5–15 %	Direction sensitivity unclear
Sentiment-augmented	3	Yes (3/3 at 7–30d)	10–15 % at medium horizon	Only at 7–30-day horizon; limited evidence
Transformer-based	4	Emerging (2/4)	12–18 % in rigorous studies	Very recent; regime coverage zero
Bayesian DL	1	Insufficient n	Calibrated intervals	Only Pantachang et al. (2022); not replicated

Note. Consistent Evidence: Yes = consistent across ≥5 independent studies; Moderate = 3–5 studies; Emerging = <3 studies. MAPE reductions are relative to standalone LSTM on same data.

5. Future Research Agenda

The directional trajectory of the cryptocurrency ML forecasting literature is clear. The field is moving from Bitcoin-only to multi-coin; from price-only to multi-modal (sentiment, on-chain, macro); from single-architecture to hybrid; and from evaluation by convention to — in a growing but still small minority — evaluation by design. Whether this trajectory produces cumulative knowledge depends entirely on whether the evaluation problem identified in Section 3.3 is resolved before the next architectural generation arrives.

Benchmark standardisation is the prerequisite for all other priorities. The design is simple: a community dataset covering Bitcoin, Ethereum, Binance Coin, and Ripple at daily frequency from January 2017 to December 2024, with four prescribed calendar-anchored splits. The training split covers 2017–2020; the validation split covers 2021 (the bull run year); the primary test split covers 2022–2023 (the FTX collapse and aftermath); and the extended test split covers 2024 (the institutional entry period). The four-split design directly addresses the single-regime evaluation failure mode: any model evaluated on all four splits has been tested across a bull run, a crash, and a recovery — the minimum market cycle diversity required for deployment-relevant claims. The technical cost is a GitHub repository and a data sharing agreement with a major exchange API. The organisational cost is adoption by 3–5 leading papers in Financial Innovation or similar venues, after which the network effects of citation and replication take over.

The community dynamics that could accelerate benchmark adoption deserve explicit modelling. The cryptocurrency forecasting literature is small enough (48 articles since 2018) that 3–4 high-visibility papers adopting a shared benchmark would constitute a qualitative majority of the literature within one publication cycle. Financial Innovation, as the target journal for this review, is positioned to catalyse this: a published call for papers explicitly requiring benchmark compliance, or an editorial standard requiring calendar-date disclosure, would reshape the incentive structure within 12 months of publication. No architectural breakthrough is required, no new data are needed, and the constraint is editorial and social rather than technical.

The regime-stratified evaluation design is the second immediate priority and the most direct response to the universal single-regime evaluation failure mode documented in Section 3.3. The design prescribes three mandatory test sub-periods: (a) the 2021 bull run (January–November 2021), when Bitcoin appreciated from USD 29,000 to USD 64,000 and Ethereum from USD 730 to USD 4,800, producing the most extreme positive price dynamics in the dataset; (b) the 2022 FTX collapse period (June–December 2022), when Bitcoin fell from USD 31,000 to USD 16,500 in six months, with the FTX bankruptcy on November 11 producing a single-day 15 % decline; and (c) the 2024–2025 institutional entry period (January 2024 onward), when Bitcoin ETF approval in January 2024 produced a new bull run to USD 73,000 in March 2024 followed by a correction. A model that maintains MAPE below 3 % across all three sub-periods is genuinely regime-agnostic. A model that performs well only in sub-period (a) or (b) but not both is regime-specific and should not be claimed as a general cryptocurrency forecasting solution. No current study establishes this claim for any architecture.

Cross-coin generalisation is the 2–5-year priority that determines the practical scope of the field. The preliminary evidence from the 5 multi-coin studies (Yasir et al., 2023; Younas et al., 2025; Han et al., 2025; Kehinde et al., 2025; Wang et al., 2025) suggests performance degradation of 30–80 % when applying Bitcoin-trained models to altcoins without retraining. Whether this reflects market microstructure differences (liquidity, order size, manipulation susceptibility) or distributional shift in the price-generating process is unknown. A deliberate transfer learning experiment — training on the top-10-capitalisation coins, evaluating zero-shot and few-shot on 20 altcoins selected to span a range of market capitalisation tiers and liquidity levels — would generate the first systematic boundary-condition evidence. The finding, whatever it is, would be informative: either the degradation is modest (validating general-purpose models) or severe (requiring coin-specific architectures for deployment).

Market microstructure integration is a 2–5-year priority requiring commercial data access. Gurgul et al. (2025) demonstrate order-book-based gains at intraday horizons in the single study that addresses this level of data granularity. No study extends microstructure data to daily or weekly horizons, where institutional forecasting typically operates. A controlled comparison of four feature sets — price-only, on-chain-augmented, sentiment-augmented, and order-book-augmented — on a shared architecture and shared benchmark dataset would generate the first systematic feature attribution evidence in this literature. The study requires commercial access to limit order book data, which is available from Binance, Coinbase, and Kraken via institutional data partnerships. The key design requirement is architectural consistency (to isolate feature effects) and benchmark alignment (to enable comparison with prior work).

Explainability is the 5-year horizon priority with the clearest regulatory pathway. Under SFDR Article 8/9 (EU Sustainable Finance Disclosure Regulation), any ML-based cryptocurrency trading signal deployed by a regulated fund manager must be attributable to specific inputs when challenged. A Transformer model applied to 40 features cannot satisfy this requirement without explicit XAI implementation. SHAP values and attention weight visualisations are available as off-the-shelf extensions to every architecture in the corpus; no paper applies them because no journal currently requires them. A shift in editorial policy at Financial Innovation — requiring SHAP reporting for any paper claiming state-of-the-art performance — would close this gap within one publication cycle without requiring any new methodological development. Pantachang et al. (2022) provide a partial template: their Bayesian DL approach produces probabilistic forecasts rather than point predictions, which is a step toward distributional explainability. The research gaps shown on Table 4 and also roadmap on Table 5.

Table 4. Research gaps with evidence and priority ratings.

Gap	Evidence	Coverage	Recommended Design	Priority
Benchmark standardisation	MAPE 1.7 %–4.8 % same architecture same target	<2 %	BTC/ETH/BNB/XRP 2017–2024; 4-split calendar-anchored	★★★★★
Regime-stratified evaluation	Zero studies test across 2021 bull + 2022 crash + 2024 ETF period	0 %	3 mandatory test sub-periods; all claims require coverage	★★★★★
Cross-coin generalisation	5 multi-coin studies; zero-shot untested	<10 %	Train BTC top-10; zero-shot + few-shot on 20 altcoins	★★★★☆
Microstructure integration	Order book in 1 study only (Gurgul et al., 2025)	2 %	4 feature sets × shared architecture; benchmark dataset required	★★★★☆
Explainability (XAI)	SHAP/attention in 2/48 articles (4.2 %)	4.2 %	Mandatory SHAP + attention for all Transformer/ensemble papers	★★★☆☆

Note. ★★★★★ = highest priority. Regime-stratified evaluation is the only gap affecting all 48 articles.

Table 5. Five-year future research roadmap structured as a dependency chain.

Horizon	Priority Area	Key Question	Recommended Design	Expected Contribution
0–2 yr	Benchmark dataset	Can 4-split calendar dataset standardise MAPE?	BTC/ETH/BNB/XRP; 4 calendar splits; open-source code	Cross-study comparison; meta-analysis enabled
0–2 yr	Regime-stratified eval.	Does DL maintain MAPE across bull, crash, recovery?	3 mandatory sub-periods: 2021/2022/2024	First regime-agnostic performance evidence
2–5 yr	Cross-coin generalisation	Does BTC-trained Transformer generalise to altcoins?	Zero-shot + few-shot: top-10 train, 20 altcoin test	Deployment scope boundaries established
2–5 yr	Microstructure integration	Which feature set adds most at daily horizon?	4 feature sets × shared architecture on benchmark	Feature attribution evidence for deployment
5+ yr	Explainability	Can SHAP/attention meet SFDR disclosure?	Mandatory SHAP + attention all Transformer papers	Regulatory compliance; institutional adoption
5+ yr	Developing-economy crypto	Do dynamics differ in remittance/hedge contexts?	ML on SSA/LatAm crypto; regime comparison	Geographic validity; policy relevance

Note. SFDR = Sustainable Finance Disclosure Regulation (EU); SSA = Sub-Saharan Africa; LatAm = Latin America. Horizon anchored to 2025.

6. Conclusion and Suggestion

This systematic review of 48 cryptocurrency ML forecasting articles (2018–2025) identified a structural evaluation problem — five failure modes affecting 39 of 48 articles — that prevents the literature from generating cumulative knowledge. The fourfold MAPE range (1.7 %–4.8 %) for architecturally equivalent LSTM models on the same Bitcoin target is not scientific disagreement; it is methodological inconsistency that no architectural innovation can resolve. Three architectural findings are established conditional on this caveat: CNN-LSTM hybrids outperform standalone LSTM consistently (9 of 12 studies), decomposition preprocessing consistently reduces errors (5 of 5 replications), and sentiment augmentation improves medium-horizon direction accuracy (3 of 3 studies). These findings are real, but they are

weaker than they appear because the evaluation conditions under which they were established are not shared across papers.

The field's trajectory is directionally correct: from Bitcoin-only to multi-coin (Yasir et al., 2023; Younas et al., 2025; Han et al., 2025; Kehinde et al., 2025), from price-only to multi-modal (Younas et al., 2025; Wang et al., 2025), from single-architecture to hybrid and Transformer-based (Pečiulis et al., 2024; Han et al., 2025). But the absence of regime-stratified evaluation means that none of the architectural advances has been tested under the market conditions — a crash followed by a recovery, or a regulatory shock followed by an institutional entry — that define the environment in which deployed models actually operate.

The regime-stratified evaluation design proposed in Section 5 addresses this directly. Three mandatory test sub-periods — the 2021 bull run, the 2022 FTX collapse, and the 2024–2025 institutional entry — are the minimum standard for deployment-relevant claims. A model that maintains MAPE below 3 % across all three is genuinely regime-agnostic and can be credibly presented as a deployment candidate. No current model has been tested to this standard.

The roadmap in Table 6 is structured as a dependency chain. Each priority depends on the prior one: cross-coin generalisation requires benchmark standardisation; microstructure integration requires cross-coin generalisation evidence; explainability requires all previous layers. The field cannot shortcut this sequence by working on explainability before establishing what, exactly, the models being explained are doing across market regimes. Benchmark standardisation and regime-stratified evaluation are prerequisites; all subsequent priorities — cross-coin generalisation, microstructure integration, explainability — require a shared evaluation foundation to produce results that can be compared across papers. A field that produced 22 articles in the first five months of 2025 has the collective capacity to establish that foundation within 24 months. Whether it does depends on whether journal editors, research funders, and leading research groups are willing to coordinate on shared standards rather than compete on private evaluation conditions.

References

Ahamed, S. A., & Ravi, C. (2021). Prediction of cryptocurrency prices using LSTM and ANN. International Journal of Swarm Intelligence and Evolutionary Computation, 10(6), 1–8. doi: 10.4018/IJSIR.2021040102

Altan, A., Karasu, S., & Bekiros, S. (2019). Digital currency forecasting with chaotic meta-heuristic signal processing. Chaos, Solitons and Fractals, 126, 325–336. doi: 10.1016/j.chaos.2019.07.011

Ammer, M. A., & Aldhyani, T. H. H. (2022). Deep learning algorithm to predict cryptocurrency fluctuation prices. Electronics, 11(15), 2349. doi: 10.3390/electronics11152349

Bengio, Y., Courville, A., & Vincent, P. (2013). Representation learning: A review and new perspectives. IEEE Transactions on Pattern Analysis and Machine Intelligence, 35(8), 1798–1828. doi: 10.1109/TPAMI.2013.50

Farooq, A., Irfan Uddin, M., Adnan, M., Alamer, A., Almutairi, S., & Ullah, Z. (2024). Bidirectional LSTM for cryptocurrency forecasting. Heliyon, 10(22), e40142. doi: 10.1016/j.heliyon.2024.e40142

Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep learning. MIT Press. doi: 10.7551/mitpress/9780262035613.001.0001

Gurgul, V., Lessmann, S., & Härdle, W. K. (2025). Deep learning for intraday cryptocurrency forecasting with order book data. International Journal of Forecasting, 41(2), 601–621. doi: 10.1016/j.ijforecast.2025.02.007

Han, P., Chen, H., Rasool, A., Jiang, Q., & Yang, L. (2025). Decomposition-Transformer hybrid for multi-step cryptocurrency price prediction. Expert Systems with Applications, 265, 125515. doi: 10.1016/j.eswa.2024.125515

Hitam, N. A., & Ismail, A. R. (2018). Comparative performance of ML algorithms for cryptocurrency forecasting. Indonesian Journal of Electrical Engineering and Computer Science, 11(3), 1121–1128. doi: 10.11591/ijeecs.v11.i3.pp1121-1128

Hong, Q. N., Pluye, P., Fabregues, S., Bartlett, G., Boardman, F., Cargo, M., Dagenais, C., Gagnon, M.-P., Griffiths, F., Nicolau, B., O'Cathain, A., Rousseau, M.-C., & Vedel, I. (2018). Mixed Methods Appraisal Tool (MMAT) version 2018. Education for Information, 34(4), 285–291. doi: 10.3233/EFI-180221

Kehinde, T. O., Adedokun, O. J., Joseph, A., Kabir, Y., & Aliyu, K. (2025). Helformer: Attention-based deep learning for cryptocurrency prediction. Journal of Big Data, 12(1), 1–28. doi: 10.1186/s40537-025-01135-4

Ladhari, A., & Boubaker, H. (2024). Fractional integration and deep learning for cryptocurrency forecasting. Forecasting, 6(2), 16. doi: 10.3390/forecast6020016

Lahmiri, S. (2020). Minute-ahead stock return forecasting based on singular spectrum analysis and support vector regression. Applied Mathematics and Computation, 370, 124951. doi: 10.1016/j.amc.2019.124951

Lahmiri, S., & Bekiros, S. (2019). Cryptocurrency forecasting with deep learning chaotic neural networks. Chaos, Solitons and Fractals, 118, 35–40. doi: 10.1016/j.chaos.2018.11.014

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. Biometrics, 33(1), 159–174. doi: 10.2307/2529310

Li, Y., Jiang, S., Li, X., & Wang, S. (2022). Hybrid data decomposition-based deep learning for Bitcoin prediction. Financial Innovation, 8, 43. doi: 10.1186/s40854-022-00336-7

Livieris, I. E., Kiriakidou, N., Stavroyiannis, S., & Pintelas, P. (2021). An advanced CNN-LSTM model for cryptocurrency forecasting. Electronics, 10(3), 287. doi: 10.3390/electronics10030287

Moher, D., Shamseer, L., Clarke, M., Ghersi, D., Liberati, A., Petticrew, M., Shekelle, P., & Stewart, L. A. (2015). Preferred reporting items for systematic review and meta-analysis protocols (PRISMA-P) 2015 statement. Systematic Reviews, 4, 1. doi: 10.1186/2046-4053-4-1

Nakamoto, S. (2008). Bitcoin: A peer-to-peer electronic cash system. Retrieved from https://bitcoin.org/bitcoin.pdf

Nazareth, N., & Reddy, Y. V. R. (2023). Financial applications of machine learning: A literature review. Expert Systems with Applications, 219, 119640. doi: 10.1016/j.eswa.2023.119640

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Gluud, C., Mayo-Wilson, E., McDonald, S., McGuinness, L. A., Stewart, L. A., Thomas, J., Tricco, A. C., Welch, V. A., … Moher, D. (2021). The PRISMA 2020 statement. BMJ, 372, n71. doi: 10.1136/bmj.n71

Pantachang, K., Tansuchat, R., & Yamaka, W. (2022). Bayesian deep learning for Bitcoin price prediction. Axioms, 11(10), 527. doi: 10.3390/axioms11100527

Pečiulis, T., Ahmad, N., Menegaki, A. N., & Biržinytė, I. (2024). Transformer model for multi-step Bitcoin price forecasting. Forecasting, 6(3), 3114. doi: 10.1002/for.3114

Rafi, M., Mirza, Q. A. K., Sohail, M. I., & Aliasghary, M. (2023). Enhancing cryptocurrency price forecasting accuracy via feature selection. IEEE Access, 11, 62940–62956. doi: 10.1109/ACCESS.2023.3287888

Sebastiao, H., & Godinho, P. (2021). Forecasting and trading cryptocurrencies with machine learning. Financial Innovation, 7, 3. doi: 10.1186/s40854-020-00217-x

Shamshad, H., Ullah, F., Ullah, A., & Kebande, V. R. (2023). Forecasting and trading stable cryptocurrencies with machine learning. IEEE Access, 11, 117369–117387. doi: 10.1109/ACCESS.2023.3327440

Syed, S., Talha, S. M., Iqbal, A., & Ahmad, N. (2024). Ensemble machine learning for Bitcoin direction prediction. AI, 5(4), 136. doi: 10.3390/ai5040136

Wang, M., Braslavski, P., & Ignatov, D. I. (2025). Temporal Fusion Transformer for multi-step cryptocurrency forecasting. Forecasting, 7(3), 48. doi: 10.3390/forecast7030048

Yasir, M., Attique, M., Latif, K., & Chaudhary, G. M. (2023). Deep learning for cryptocurrency forecasting using social media sentiment. Journal of Enterprise Information Management, 36(2), 522–547. doi: 10.1108/JEIM-02-2020-0077

Younas, R., Raza Ur Rehman, H. M., & Choi, G. S. (2025). CNN-GRU with sentiment features for multi-coin cryptocurrency prediction. Journal of Big Data, 12(1), 91. doi: 10.1186/s40537-025-01291-7