Enigma in Economics
https://enigma.or.id/index.php/economy
e-ISSN: 3026-6696

Can Technical Indicators Predict Daily Price Direction? Walk-Forward Evidence from Bitcoin, the S&P 500, and Gold

Ifah Shandy[1*], Benyamin Wongso[2]

[1]*Department of Mathematics Education, Enigma Institute, Palembang, Indonesia*
[2]*Department of Social Science Education, Enigma Institute, Palembang, Indonesia*

A R T I C L E I N F O

Article history:
Received: May 25, 2026
Accepted: June 29, 2026
Published: July 30, 2026

Keywords:
Bitcoin
Gold
S&P 500
Technical analysis
Walk-forward validation

***Corresponding author:**
Ifah Shandy
ifah.shandy@enigma.or.id

ORCID:
Ifah Shandy: 0009-0002-1705-0840
Benyamin Wongso: 0009-0001-4047-8581

DOI:
https://doi.org/10.61996/economy.v4i1.132

All authors have reviewed and approved the final version of the manuscript.

A B S T R A C T

Background. Technical analysis remains among the most accessible forms of market decision support, but its incremental predictive value depends on design choices, and repeated model selection can create spurious backtest performance.

Objective. This study evaluates whether widely used technical indicators, reconstructed from public formulas, can predict the next daily price direction of Bitcoin, the S&P 500, and gold.

Methods. Daily data from 1 January 2015 to 14 July 2026 are examined using a chronological walk-forward design. Every signal observed at the close of day t is matched only with the sign of the close-to-close return from t to t+1. The main out-of-sample period begins in 2019, with 2019–2022 used for model selection and 2023–2026 reserved for confirmation. Performance is measured primarily by balanced accuracy and supplemented by accuracy, directional recall, stationary-bootstrap confidence intervals, and after-cost trading outcomes.

Results. The best single indicator, Ichimoku 9/26/52, produced a macro balanced accuracy of 50.9%, while the best predetermined combination, Volume Confirmed, reached 50.8%. A new ridge-logistic hybrid indicator, SPAH-1, achieved 51.9% in validation and 51.3% in confirmation. Its confirmation balanced accuracy was 49.8% for Bitcoin, 49.2% for the S&P 500 proxy, and 55.0% for gold; only gold's bootstrap interval excluded 50%, but its predictions were strongly biased toward the upward class. After trading costs, Volume Confirmed underperformed buy-and-hold for all three assets. Additional selective experiments did not support an 80% daily prediction target.

Conclusion. Technical indicators may assist regime description and decision confirmation, but they do not provide a robust universal next-day forecasting edge.

1. Introduction

Technical analysis remains one of the most accessible forms of market decision support. Moving averages, oscillators, trend filters, and volume-based measures convert price and volume histories into reproducible signals. Recent stock-market studies show that technical indicators remain common inputs for direction classification, although their incremental value depends strongly on feature selection, model design, and the evaluation period[1–3]. In this article, the indicators are treated as transparent mathematical transformations with fixed parameters rather than as features of any particular charting product. The practical attraction is clear: if the completed daily bar contains information about the next daily move, a user can translate it into a rule without estimating a structural valuation model. Accessibility, however, is not predictive validity; a chart may describe the current state accurately while adding little information about the sign of the next return.

A stringent test is required because recent evidence remains mixed. Large-scale studies of technical rules find that apparent profitability can weaken after false-discovery control, while risk adjustment and trading frictions can materially alter the economic interpretation[4,5]. The unresolved question is therefore not whether one historical configuration can look successful, but whether recognizable indicator families retain class-balanced next-day information across dissimilar assets after a common timing rule, chronological selection, a frozen confirmation period, and explicit economic frictions.

Bitcoin, broad United States equities, and gold provide a demanding cross-asset laboratory. Research on Bitcoin has combined technical indicators, machine learning, on-chain variables, macroeconomic conditions, and regime-dependent models, with performance varying across samples and feature sets[6–10]. Broader cryptocurrency studies likewise document distinctive return drivers, changing market conditions, and difficult volatility dynamics[11–15]. Gold adds a

monetary and defensive asset whose price dynamics have recently been examined with machine-learning and reconfigured long-memory features[16,17]. The S&P 500 represents a mature equity market with deep liquidity and a pronounced positive drift. A rule that works across all three assets would be more persuasive than one that succeeds only in one market or regime; disagreement can instead reveal asset-specific structure.

Classification accuracy alone is not sufficient evidence. A model that predicts increases almost all the time may appear accurate in a rising market while rarely identifying declines. Balanced accuracy gives equal weight to upward and downward recall and is therefore more informative when class proportions differ. Serial dependence is another concern because daily observations and overlapping horizons are not independent. Predictive and economic evaluation must consequently report class-specific outcomes, dependence-aware uncertainty, turnover, and trading costs together.

Machine-learning methods can combine many weak predictors, but they expand the opportunity for overfitting. Recent asset-pricing evidence shows that nonlinear models can enhance anomaly signals, yet controlled studies demonstrate that repeated model choice can create spurious backtest performance[18–21]. Chronological validation is therefore essential; research on cryptocurrency forecasting also shows that walk-forward design choices affect the credibility of reported accuracy[22]. Regularization, frozen confirmation windows, transparent accounting for rejected alternatives, and a complete prediction archive are necessary safeguards. Even with these controls, a result only modestly above 50% should be interpreted cautiously rather than marketed as a dependable trading system.

The present study addresses four questions. First, when a technical signal is formed at the close of day t, how accurately does it predict whether the next available daily close will be higher or lower? Second, do predetermined combinations of trend, momentum, and volume indicators improve on their components? Third, can a regularized cross-asset model, the Statistical Predictive Adaptive Hybrid version 1 (SPAH-1), deliver a stable improvement in a confirmation period that was not used for parameter selection? Fourth, is an aspirational 80% directional accuracy attainable when the system may abstain on low-confidence observations? These questions are evaluated using Bitcoin, the SPDR S&P 500 ETF Trust (SPY) as a tradable S&P 500 proxy, and continuous gold futures.

The novelty lies in combining four safeguards that are often examined separately: a bar-by-bar decision timestamp, one harmonized framework across cryptoassets, equities, and gold, a model architecture frozen before an untouched confirmation window, and an explicit falsification test of the 80% claim using coverage and class-balanced metrics. The article connects each daily signal with the following close-to-close direction; compares recognized indicators, predetermined combinations, parameter sensitivity, market regimes, and a regularized model; and reports negative findings, confidence intervals, class-specific recall, trading costs, and selective coverage. The purpose is to estimate how much next-day information transparent, formula-based rules retain after conservative confirmation procedures.

Three expectations guide interpretation rather than serve as post hoc success criteria. First, a genuinely portable indicator should remain above chance across more than one asset instead of relying on a single favorable market. Second, a selected parameter should preserve its ordering after the 2022 freeze; a large validation gain followed by confirmation reversal is evidence of instability, not adaptation. Third, an economically useful classifier should identify both directions with non-trivial coverage and should not obtain its headline accuracy by predicting only the dominant class. These expectations connect statistical discrimination, confirmation stability, and implementation relevance. They also explain why the analysis reports negative results and rejected candidates: a complete record is necessary to distinguish a repeatable signal from a visually compelling configuration selected after observing the outcomes.

2. Methods

2.1 Research design and data

The study used daily open, high, low, close, and volume observations covering 1 January 2015 through 14 May 2026, subject to each market's native trading calendar. Bitcoin was represented by the Yahoo Finance BTC-USD aggregate series, the S&P 500 by the SPDR S&P 500 ETF Trust ticker SPY, and gold by Yahoo Finance continuous COMEX gold futures ticker GC=F. These vendor-specific feeds are not interchangeable with every exchange, venue, or continuous-contract construction. Daily dates were retained from the vendor feed and were not realigned to a single global closing timestamp; each prediction therefore refers to the next available daily bar for that series. The feed mapping, coverage dates, and observation counts are detailed in Table 1. Bitcoin contributes 4,213 bars, compared with 2,898 for SPY and 2,903 for gold, because Bitcoin trades seven days per week while the other assets follow exchange calendars.

Table 1. Data coverage and symbol mapping. Data through 14 July 2026.

Asset	Study feed	Comparable market representation	Coverage	Bars	Notes
BTC	BTC-USD	BTCUSD / BTCUSDT	2015-01-01 to 2026-07-14	4,213	24/7; Yahoo aggregated volume
SPY	SPY	AMEX:SPY	2015-01-02 to 2026-07-14	2,898	S&P 500 proxy with ETF volume
Gold	GC=F	COMEX:GC1!	2015-01-02 to 2026-07-14	2,903	Continuous futures; rolling may create gaps

Source: Authors' calculations from the study dataset.

The dataset was arranged chronologically and each asset retained its native sequence of available daily bars; closed-market dates were not paired artificially across assets. An asset-day became eligible only after all inputs required by the applicable indicator or model were non-missing and every rolling window had completed its warm-up. Consequently, denominators can differ across rules and are reported with the results. The principal out-of-sample evaluation began on 1 January 2019. Observations from 2019 through 2022 formed the validation window for parameter and architecture selection. All final model choices, regularization strengths, transformations, and decision thresholds were fixed at the end of 2022. Observations from 1 January 2023 through 14 May 2026 then served as a confirmation window. No confirmation observation was used to select an architecture or tune a parameter. Validation results therefore describe selection performance, whereas confirmation results approximate post-freeze performance.

Public secondary market series were used and the study involved no human participants, interventions, private personal data, or animal subjects. Institutional ethics approval was therefore not required. The analysis is educational and does not constitute investment advice. Vendor-specific choices remain material: SPY histories may differ according to split, dividend, and adjusted-close treatment, while GC=F depends on the vendor's continuous-contract construction, rollover convention, and later revisions. The source analysis did not independently reconstruct those vendor transformations from exchange-level contracts. Exact numerical replication therefore requires the archived data snapshot, acquisition date, adjustment fields, software environment, parameter grid, random seed, and daily prediction file. Symbols, sample windows, formulas, frozen coefficients, and the bootstrap seed are recorded here to support reconstruction, but the vendor snapshot and executable replication package should accompany submission.

2.2 Target definition and timing controls

For every eligible asset-day t, the target compared the next available close with the current close. Zero returns were excluded from directional denominators. A positive signal issued only after the complete daily bar at t was available was correct only if the next close was higher; a negative signal was correct only if it was lower. All rolling features, standardization inputs, and regime labels used information available no later than bar t. Future bars, future-filled missing values, and confirmation-period observations were not permitted in the day-t decision. The formal target definition is stated below.

$$y_{t+1} = 1 \text{ if } C_{t+1} > C_t;\ y_{t+1} = 0 \text{ if } C_{t+1} < C_t.$$

The primary horizon was one daily bar. Longer horizons were examined only as robustness or addendum analyses. For a k-bar horizon, the outcome compared Close(t+k) with Close(t). Overlapping horizons can exaggerate the apparent sample size because adjacent targets share returns. Full-coverage horizon summaries retained overlapping observations for descriptive comparability, whereas the selective addendum selected and scored candidates on non-overlapping validation and confirmation observations. The main one-day analysis used all eligible consecutive observations and applied time-series-aware uncertainty estimation. These two estimands are reported separately and should not be compared as though their effective sample sizes were identical.

2.3 Technical indicators and directional rules

Fifteen default indicator families were reconstructed from price and volume inputs: exponential moving-average price filters, an EMA10/EMA200 crossover, MACD, ADX with directional movement, Supertrend, Ichimoku 9/26/52, RSI, Stochastic RSI, DSS Bressert, OBV, MFI, VWMA, a rolling volume-profile point of control, a composite technical rating, and a composite regime indicator. Recent technical-indicator studies informed the reproducible feature definitions and directional translations[1–3,6,7]. The exact parameters and decision rules used in this study are detailed in Table 2.

Table 2. Default parameters and tested directional rules.

Indicator	Default parameters	Directional rule	Function
EMA price	10 and 200	Long when Close > EMA; short when below	Fast/slow trend
EMA cross	10/200	Long when EMA10 > EMA200	Regime trend
RSI	14	Long when RSI > 50; short when RSI < 50	Momentum regime
Stoch RSI	14/14/3/3	Long when %K > %D	Timing oscillator
DSS Bressert	10/3/3	Double-smoothed stochastic; long when DSS > signal	Timing oscillator
MACD	12/26/9	Long when histogram > 0	Momentum trend
DMI/ADX	14	Direction from +DI versus -DI only when ADX >= 20	Direction and strength
Bollinger %B	20, 2 SD	Long when %B > 0.5	Position within band
ROC	20	Long when ROC > 0	Price momentum
MFI	14	Long when MFI > 50	Price and volume
OBV	EMA 20 OBV	Long when OBV > EMA20(OBV)	Volume accumulation
Volume Profile	60 bar, 24 bins	Long when Close > rolling POC	Acceptance price
VWMA	20	Long when Close > VWMA20	Volume-weighted price
Supertrend	ATR 10, multiplier 3	Direction of the Supertrend line	Trend and volatility
Ichimoku	9/26/52	Long above cloud; short below; neutral inside	Regime trend

Source: Authors' calculations from the study dataset.

Continuous indicator values were translated into directional states by rules specified before comparing final results. Trend filters generally assigned +1 when price or a faster component exceeded a slower reference and -1 otherwise. Oscillator rules used relative positions, directional crossings, or neutral zones as appropriate. Volume measures were interpreted through their slopes, price relationships, or agreement with price direction. These translations are deliberately simple because the research question concerns how recognizable chart signals behave when converted into reproducible daily decisions. They should not be interpreted as the only possible implementation of each indicator.

Seven predetermined combinations tested whether confirmation across families improved reliability. Trend Core aggregated EMA200, EMA10/200, MACD, and Supertrend; Momentum Core combined RSI, Stochastic RSI, and DSS; and Volume Confirmed combined the volume-profile signal with VWMA, OBV, and MFI evidence. Other combinations joined trend and momentum, trend and volume, all available votes, or DSS/Stochastic-RSI/EMA200. Majority voting was used unless a more specific condition is detailed in Table 3. Ties or neutral states were omitted where the rule did not produce a direction.

Table 3. Indicator combinations specified before comparison of outcomes.

Combination	Component	Decision
Trend Core	EMA200, EMA10/200, MACD, Supertrend	Majority sign
DSS + Stoch RSI + EMA200	DSS, Stoch RSI, EMA200	Majority sign
Trend + Momentum + ADX	EMA10/200, RSI, MACD, DMI/ADX, Supertrend	Majority sign
Volume Confirmed	EMA10/200, POC, VWMA, OBV, MFI	>=3 of 5 bullish = long
Balanced Ensemble	EMA200, RSI, DSS, MACD, DMI/ADX, POC, Supertrend	Majority sign
Regime Adaptive	Trend Core when ADX >= 20; oscillator vote when ADX < 20	Switch by regime
Walk-forward Ridge Logit	16 continuous features; trailing 1,000 bars	Probability >= 50% = long

Source: Authors' calculations from the study dataset.

2.4 Performance measures and inference

The primary measure was balanced accuracy (BAcc), supported by ordinary accuracy, upward recall, downward recall, prediction count (N), the predicted-upward share, and, where applicable, executed-signal coverage. BAcc gives equal weight to the two outcome classes and is suitable when class proportions differ[23]. BAcc of 50% represents chance-level class-balanced discrimination. Confusion-matrix counts were audited by asset and rule so that UP, DOWN, NEUTRAL, executed, and scored denominators could be reconciled. The metric definitions are stated below.

$$BAcc = 1/2(TPR + TNR),\ TPR = TP/(TP + FN),\ TNR = TN/(TN + FP).$$

Uncertainty around BAcc was estimated with the stationary bootstrap[24], using circular blocks averaging

20 daily bars, 600 replications for indicator comparisons, 1,200 replications for SPAH-1, and random seed 20260715. The reported 95% intervals are conditional on the feed, block design, historical process, and implementation; they are not adjusted for the full family of indicators, parameters, architectures, horizons, and thresholds examined. The frozen confirmation window reduces selection leakage but does not eliminate multiplicity. Because no formal familywise correction was applied, best-rule rankings and the gold finding remain descriptive. Recent controlled backtest studies reinforce the need for this restraint[19–21].

For economic interpretation, directional signals were mapped to unit long or unit short exposure; neutral observations in selective experiments implied no new executed signal. The source analysis evaluated compound annual growth rate, annualized Sharpe ratio, maximum drawdown, turnover, and a simple transaction-cost deduction against buy-and-hold. Classification scoring is independent of fill assumptions, but economic returns are not. Because the retained source outputs do not contain exchange-specific executable quotes or a complete trade ledger, the simulation cannot establish a tradable close-t fill and does not fully model bid-ask spread, slippage, commissions, financing, borrow availability, leverage or margin, taxes, and futures roll costs. The reported after-cost outcomes are therefore a stylized feasibility and opportunity-cost check, not implementable net returns. Buy-and-hold is retained as the economic baseline because frequent long/short switching sacrifices persistent long exposure as well as paying direct costs.

2.5 SPAH-1 model development

SPAH-1 tested whether a disciplined statistical combination could extract information that simple voting discarded. The feature set included trend, momentum, volume, volatility, and regime variables available by close t. Ridge-penalized logistic regression was chosen for transparent probabilities and coefficient shrinkage in correlated feature sets. The walk-forward sequence was: define chronological training and validation rows; estimate means and standard deviations on training rows only; standardize and clip subsequent values to [-6,+6]; fit every permitted ridge candidate; choose architecture, penalty, and cutoff using 2019-2022 validation only; refit through 31 December 2022; freeze coefficients and transformations; and generate 2023-2026 confirmation predictions once, without retuning.

Five candidate architectures were compared in validation: a pooled cross-asset model, a pooled regime model, asset-specific models, a regime-plus-asset structure, and a hybrid that averaged pooled-regime and asset-specific probabilities. Regularization strength and probability cutoffs were chosen exclusively from the 2019–2022 validation results. The hybrid architecture with ridge parameter 100 and an upward cutoff of 51% was designated SPAH-1. After selection, the component models were refitted using only data available through 31 December 2022, coefficients and preprocessing statistics were frozen, and confirmation predictions were generated for 2023–2026 without further tuning. Appendix D reports every frozen coefficient, training mean, and standard deviation. Chance-level BAcc and simple always-UP behavior provide statistical baselines; buy-and-hold provides the economic baseline.

For each component c, frozen training statistics standardized feature j at time t. The component probability was obtained from a ridge-logistic score, and SPAH-1 averaged the applicable pooled-regime and asset-specific probabilities. It predicted UP when the combined probability met the frozen 51% cutoff and DOWN otherwise. The equations below use an asterisk for a standardized feature; intercepts are not standardized.

$$x^*_{j,t} = \mathrm{clip}[(x_{j,t} - \mu_{j,\mathrm{train}}) / \sigma_{j,\mathrm{train}}, -6, +6].$$

$$p_{c,t} = [1 + \exp(-z_{c,t})]^{-1}, z_{c,t} = \beta_{0,c} + \Sigma_j \beta_{j,c} x^*_{j,t}.$$

$$p_{\mathrm{SPAH\text{-}1},t} = 1/2\ [p_{\mathrm{pooled\text{-}regime},t} + p_{\mathrm{asset},t}].$$

2.6 Tests of the 80% target

An additional experiment investigated whether accuracy could approach 80% by changing the forecast horizon and permitting abstention. Full-coverage pooled and hybrid models were evaluated over one-, three-, five-, ten-, and twenty-bar horizons. Selective candidates issued UP only above a high probability threshold and DOWN only below a low threshold; observations between the thresholds were labeled neutral and excluded from executed-signal accuracy. Candidate horizons and thresholds were selected using non-overlapping validation observations and then frozen for non-overlapping confirmation. Wilson intervals were reported for executed-signal accuracy because selective samples could be small.

This addendum was designed as a falsification and exploratory stress test rather than a replacement for the confirmatory daily analysis. Selective accuracy can be increased mechanically by making fewer predictions, especially if thresholds are chosen after inspecting many alternatives. Each selective model is therefore treated as a three-state process: UP, DOWN, and NEUTRAL. Executed-signal accuracy excludes NEUTRAL observations, while coverage equals executed decisions divided by all eligible decision dates; selective risk is the error rate among executed decisions, and opportunity cost includes the outcomes forgone during abstention. Calibration was not established from an archived probability-by-outcome file and remains a required prospective diagnostic. High accuracy based on a few mostly upward events does not establish balanced directional skill. Accordingly, N, UP/DOWN composition, coverage, BAcc, and confirmation intervals are reported together. A credible claim near 80% would require a

sufficiently large forward sample, useful coverage, both directional classes, and an interval that remains convincingly above a predeclared benchmark after multiple-testing control.

3. Results

3.1 Direct audit of daily signal matching

The timing audit confirmed that each decision used the signal at close t and was scored against the direction from close t to the next available close. Representative confusion-matrix outcomes are detailed in Table 4. Sample sizes differ because Bitcoin trades every day and because indicators have different warm-up requirements or neutral states. The table also shows why ordinary and balanced accuracy can diverge: several rules favored the majority upward class, whereas others generated more symmetric decisions. No main result was produced by matching a signal with the same bar used to calculate it.

Table 4. Audit of daily direction matching, out-of-sample 2019–14 July 2026.

Asset	Indicator	N	Accuracy	BAcc	Up recall	Down recall
BTC	EMA 10 (price)	2,751	49.7%	49.7%	53.1%	46.2%
BTC	EMA 200 (price)	2,751	50.6%	50.4%	61.2%	39.6%
BTC	DSS Bressert 10/3/3	2,751	48.5%	48.4%	48.9%	48.0%
BTC	Stoch RSI 14/14/3/3	2,697	49.2%	49.2%	49.7%	48.6%
BTC	Volume Profile POC 60/24	2,751	51.2%	51.1%	56.0%	46.3%
BTC	Volume Confirmed	2,751	50.6%	50.5%	55.1%	45.9%
SPY	EMA 10 (price)	1,889	52.6%	50.8%	67.5%	34.0%
SPY	EMA 200 (price)	1,889	54.7%	51.3%	82.0%	20.6%
SPY	DSS Bressert 10/3/3	1,889	50.1%	49.6%	53.9%	45.4%
SPY	Stoch RSI 14/14/3/3	1,843	50.7%	50.3%	54.0%	46.5%
SPY	Volume Profile POC 60/24	1,889	52.9%	50.5%	72.5%	28.5%
SPY	Volume Confirmed	1,889	53.5%	51.2%	71.6%	30.8%
Gold	EMA 10 (price)	1,886	49.8%	48.8%	58.3%	39.2%
Gold	EMA 200 (price)	1,886	53.4%	50.1%	81.9%	18.2%
Gold	DSS Bressert 10/3/3	1,886	49.9%	49.8%	51.0%	48.6%
Gold	Stoch RSI 14/14/3/3	1,846	50.3%	50.2%	51.7%	48.6%
Gold	Volume Profile POC 60/24	1,886	52.0%	50.7%	63.3%	38.0%
Gold	Volume Confirmed	1,886	52.9%	50.7%	71.6%	29.7%

Source: Authors' calculations from the study dataset.

This finding matters because small timing errors could overwhelm the observed effects. Most balanced accuracies in the study are within roughly one percentage point of 50%. If a future close, future high, or revised rolling value leaked into the feature set, even a minor bias could create an apparently strong result. The explicit audit shows that the weak results are not caused by an overly harsh same-day definition; they reflect genuine next-close tests. It also makes the analysis comparable across indicators whose visual plots may otherwise suggest different entry times.

3.2 Single-indicator performance

The complete cross-asset pattern is visualized in Figure 1, while the asset-level winners and confidence intervals are detailed in Table 5. Indicators cluster tightly around chance. Ichimoku 9/26/52 has the highest macro BAcc at 50.9%, followed by Volume Profile POC 60/24 at 50.8% and EMA10/200 at 50.7%. At the asset level, the winners are Volume Profile POC for Bitcoin (51.1%), Bollinger %B 20/2 for SPY (51.3%), and Ichimoku 9/26/52 for gold (51.3%); every corresponding 95% interval includes 50%.

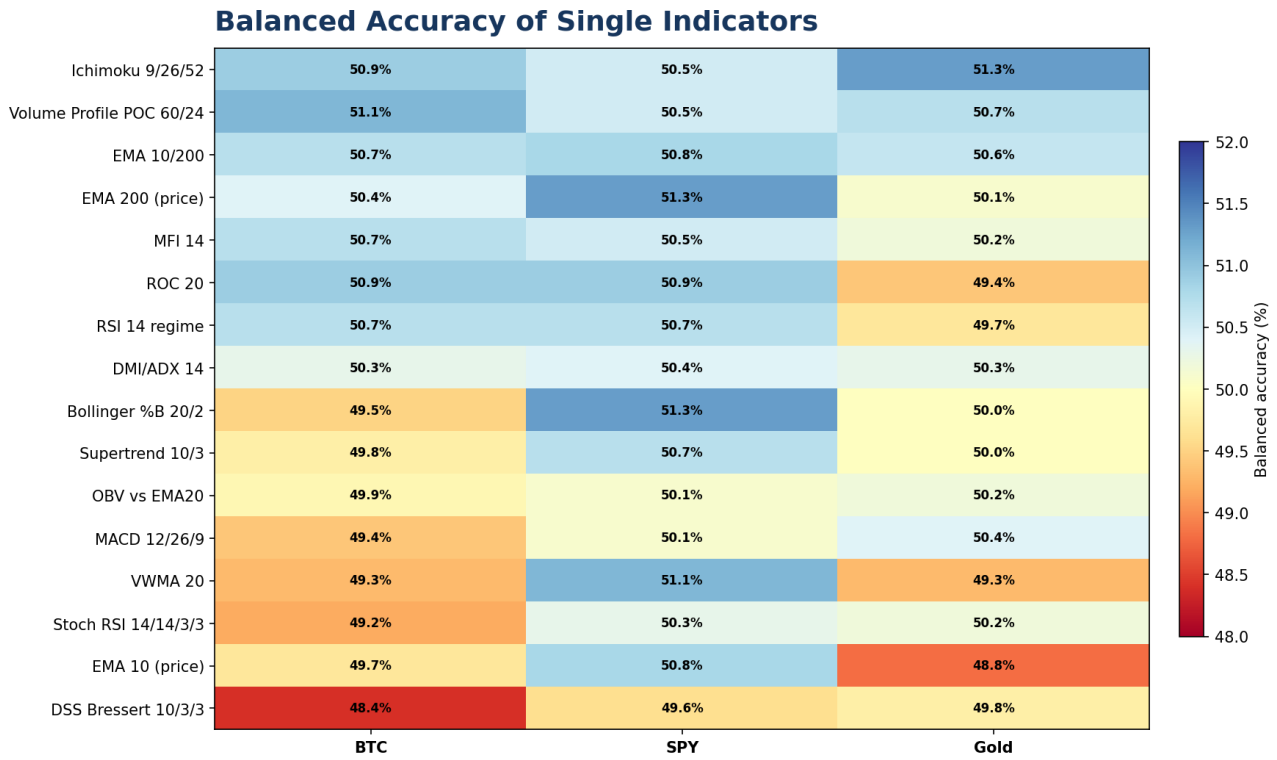

Figure 1. Balanced accuracy of single indicators by asset.
Source: Authors' calculations.

Table 5. Best single indicator by asset; all confidence intervals include 50%.

Asset	Best indicator	BAcc	95% CI	Accuracy	Sharpe	Max DD
BTC	Volume Profile POC 60/24	51.1%	49.3–52.7%	51.2%	0.88	-59.4%
SPY	Bollinger %B 20/2	51.3%	49.3–53.1%	53.4%	0.16	-35.3%
Gold	Ichimoku 9/26/52	51.3%	49.2–53.5%	53.1%	0.44	-43.9%

Source: Authors' calculations from the study dataset.

Bitcoin's best single result is therefore 51.1% BAcc, while the best SPY and gold results are both 51.3%. These exact values agree across Figure 1 and Table 5. The narrow spread is consistent with indicators that summarize recent market state but contribute little incremental information about the next daily sign. Reporting only the maximum from a large menu would be misleading because one or more approximately random rules can rank above 50% by chance.

The limited success of oscillators is economically plausible. RSI, Stochastic RSI, and DSS transform recent price changes into bounded measures of relative strength or position. Such transformations may identify extended conditions, but an extended condition can precede either continuation or reversal. A deterministic rule that always treats an overbought reading as bearish, or a crossing as immediate confirmation, ignores regime, volatility, and news. Trend indicators face the opposite problem: they can participate in persistent movement, yet their lag often makes them weak predictors of the very next close. The empirical concentration around 50% reflects these offsetting mechanisms.

3.3 Indicator combinations and economic performance

Combining indicators did not produce a material breakthrough. Figure 2 visualizes each combination's macro BAcc relative to the 50% benchmark. Volume Confirmed ranks first at 50.8%, only 0.8 percentage points above chance; Regime Adaptive follows at 50.7%, while DSS + Stoch RSI + EMA200 is 49.7%. The exact combination definitions are detailed in Table 3, and the full ranking is reported later in Table 15.

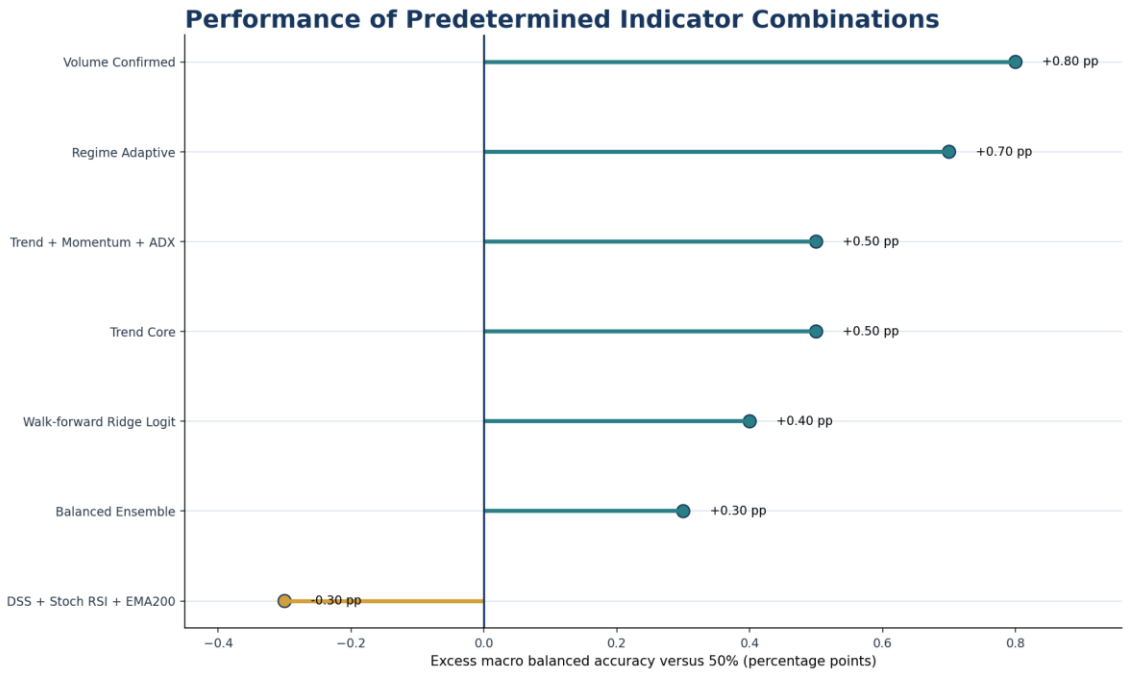

Figure 2. Excess balanced accuracy of indicator combinations relative to 50%.
Source: Authors' calculations.

Economic outcomes are detailed in Table 6 and visualized in Figure 3. After transaction costs, Volume Confirmed underperforms buy-and-hold for every asset: CAGR is 32.3% versus 45.5% for Bitcoin, 1.5% versus 15.8% for SPY, and 6.4% versus 16.6% for gold. Bitcoin's maximum drawdown improves to -53.6%, but the benefit comes with high turnover and lost long exposure during major advances. The comparison demonstrates why a small classification edge cannot be equated with attractive investment performance.

Table 6. Volume Confirmed by asset compared with buy-and-hold after costs.

Asset	BAcc	95% CI	Strategy CAGR	Sharpe	Max DD	CAGR B&H
BTC	50.5%	48.8–52.0%	32.3%	0.76	-53.6%	45.5%
SPY	51.2%	49.2–53.2%	1.5%	0.17	-41.0%	15.8%
Gold	50.7%	48.5–52.5%	6.4%	0.43	-36.4%	16.6%

Source: Authors' calculations from the study dataset.

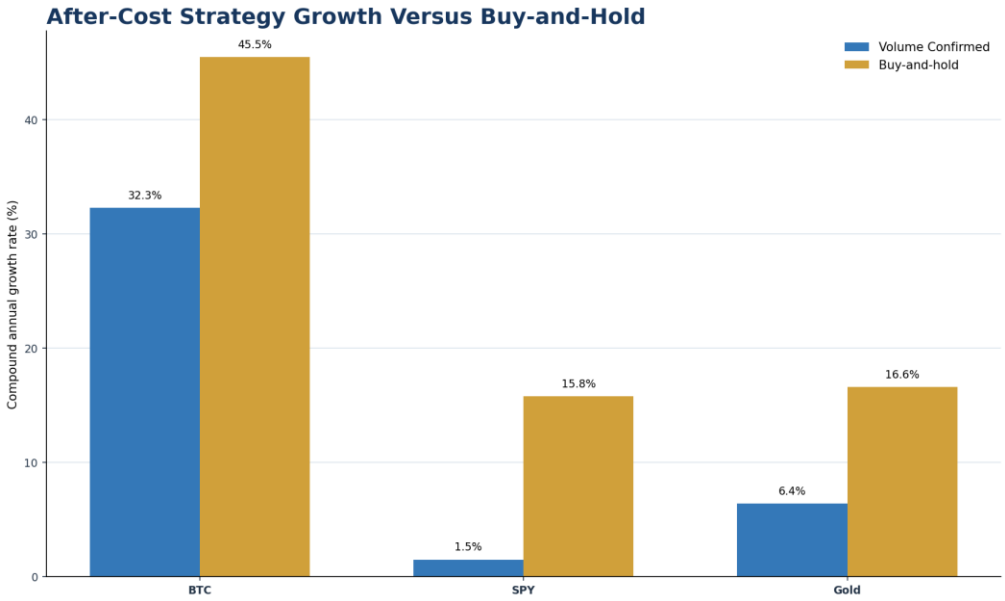

Figure 3. After-cost compound annual growth rates for Volume Confirmed and buy-and-hold.
Source: Authors' calculations.

The outcome is consistent with recent evidence that technical-rule profitability is sensitive to false-discovery control, risk adjustment, and market conditions[4,5]. It also agrees with controlled backtest research showing that repeated selection can generate spurious performance[19–21]. The present study does not prove that every technical strategy is ineffective; it shows that these transparent daily rules did not generate a robust universal next-day edge in the specified out-of-sample windows.

3.4 Parameter sensitivity and market regimes

Validation-to-confirmation changes are visualized in Figure 4 and detailed in Table 7. Several EMA, RSI, Supertrend, and DSS settings that appeared superior in 2019-2022 reverted toward or below 50% after 2023. For example, the SPY EMA 50/100 setting moves from +2.7 percentage points above chance in validation to -2.4 points in confirmation. Replacing a failed parameter with a newly optimized setting would restart the selection process and invalidate the original confirmation claim.

Figure 4. Validation and confirmation performance of selected parameters.
Source: Authors' calculations.

Table 7. Parameters selected in 2019–2022 and confirmed in 2023–14 July 2026.

Asset	Family	Selected parameter	Validation vs 50%	Confirmation vs 50%	Confirmation 95% CI
BTC	DSS Bressert	21/3/3	-1.4 pp	-1.8 pp	45.7–50.7%
SPY	DSS Bressert	14/5/3	+0.5 pp	-1.3 pp	45.6–51.8%
Gold	DSS Bressert	21/3/3	+0.1 pp	+0.7 pp	46.8–54.2%
BTC	EMA cross	5/100	+2.5 pp	-0.3 pp	47.6–52.0%
SPY	EMA cross	50/100	+2.7 pp	-2.4 pp	45.1–49.7%
Gold	EMA cross	20/200	+2.1 pp	-0.1 pp	49.0–50.7%
BTC	RSI regime	n=21	+1.8 pp	-0.2 pp	47.3–52.2%
SPY	RSI regime	n=14	+1.2 pp	-0.1 pp	47.6–52.3%
Gold	RSI regime	n=21	-0.7 pp	+0.8 pp	47.4–53.7%
BTC	Stoch RSI	21/7/3/3	-0.5 pp	+0.7 pp	48.3–53.2%
SPY	Stoch RSI	7/7/3/3	+1.2 pp	+1.2 pp	48.3–54.2%
Gold	Stoch RSI	7/7/3/3	+0.0 pp	-1.6 pp	44.8–51.7%
BTC	Volume Profile	window=120, bins=48	+3.3 pp	+0.8 pp	48.5–52.9%
SPY	Volume Profile	window=30, bins=48	+2.4 pp	+0.3 pp	47.5–52.8%
Gold	Volume Profile	window=120, bins=12	+0.7 pp	+1.8 pp	49.5–54.5%
BTC	Supertrend	14/4	+2.4 pp	-0.7 pp	46.9–51.7%
SPY	Supertrend	10/2	+1.9 pp	-1.0 pp	45.6–52.5%
Gold	Supertrend	14/4	+0.3 pp	+0.7 pp	47.6–53.6%

Source: Authors' calculations from the study dataset.

Regime-conditioned results are detailed in Table 8. Volume Confirmed remains close to chance in bull (50.6%), bear (50.6%), high-volatility (50.2%), and low-volatility (51.6%) states. None of the regime partitions supports a universal rule with a stable margin over chance. Smaller regime samples also widen uncertainty, limiting strong conclusions from local peaks.

Table 8. Average macro balanced accuracy by market regime.

Model	Bull	Bear	High volatility	Low volatility
Volume Confirmed	50.6%	50.6%	50.2%	51.6%
Volume Profile POC 60/24	50.8%	49.9%	50.2%	51.6%
Ichimoku 9/26/52	50.4%	50.7%	50.5%	51.5%
EMA 10/200	50.3%	49.6%	51.1%	50.6%
RSI 14 regime	49.7%	51.4%	50.3%	50.7%
DSS Bressert 10/3/3	49.0%	50.1%	48.6%	49.8%

Source: Authors' calculations from the study dataset.

Cross-asset differences are economically informative. Bitcoin's continuous trading, structural change, and large tail moves can destabilize daily relationships, consistent with the heterogeneous cryptocurrency evidence[6–15]. SPY has a strong long-run upward drift, so ordinary accuracy can rise simply by favoring UP predictions. Recent gold-forecasting studies also document feature and regime dependence[16,17]. The stronger SPAH-1 gold result may therefore represent a local configuration rather than a universal relationship.

3.5 SPAH-1 validation and confirmation

The five candidate architectures are compared visually in Figure 5 and numerically in Tables 9 and 10. The selected hybrid pooled/asset model reaches 51.9% macro BAcc in 2019-2022 validation and 51.3% after freezing for 2023-14 July 2026 confirmation. Table 9 details the confirmation decomposition: Bitcoin records 49.8% (N=1,285), SPY 49.2% (N=883), and gold 55.0% (N=848). Only the gold interval, 52.0%-58.2%, excludes 50%.

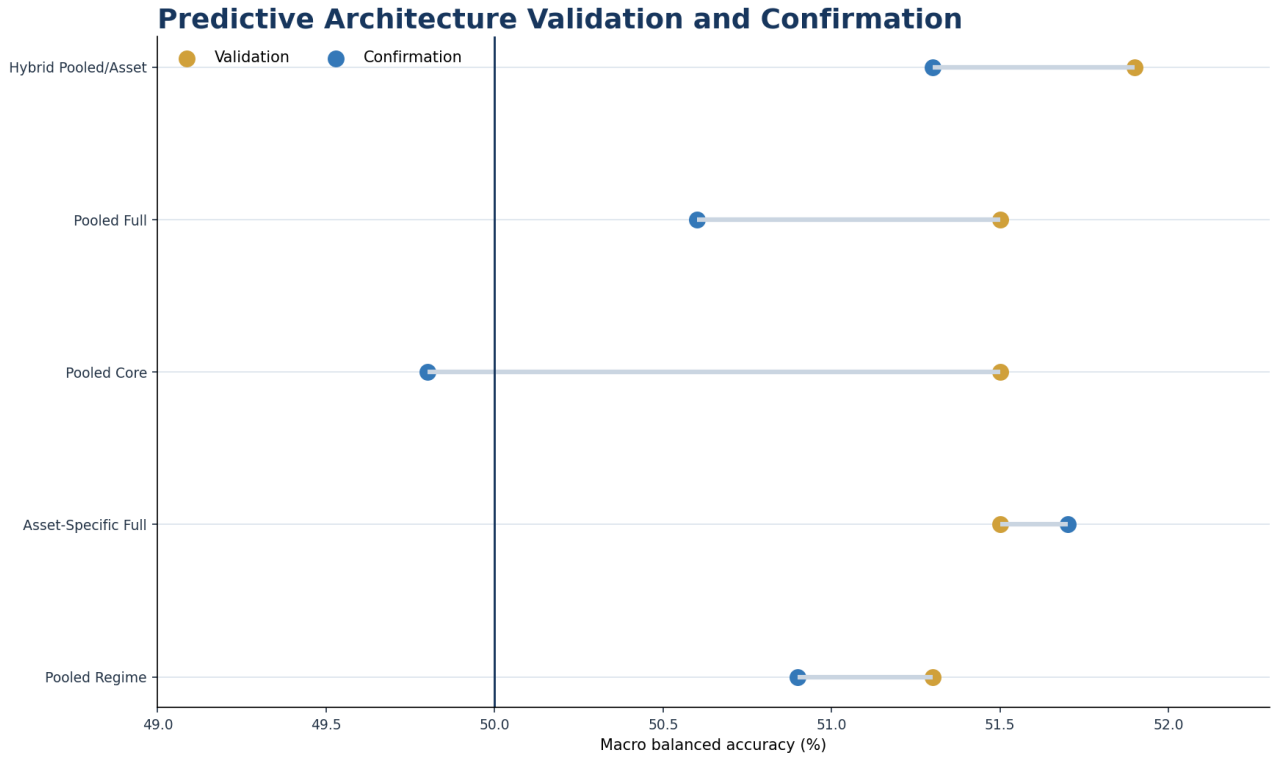

Figure 5. Validation and confirmation of five predictive architectures.
Source: Authors' calculations.

Table 9. SPAH-1 performance in the confirmation window, 2023–14 July 2026.

Asset	N	Accuracy	BAcc	95% CI BAcc	Up recall	Down recall	Predicted up
BTC	1,285	49.9%	49.8%	47.7–51.7%	80.0%	19.7%	80.2%
SPY	883	53.9%	49.2%	47.0–51.4%	84.0%	14.4%	84.7%
Gold	848	57.1%	55.0%	52.0–58.2%	76.4%	33.5%	71.9%

Source: Authors' calculations from the study dataset.

Figure 6 makes the class imbalance of SPAH-1 explicit. Upward versus downward recall is 80.0% versus 19.7% for Bitcoin, 84.0% versus 14.4% for SPY, and 76.4% versus 33.5% for gold, exactly matching Table 9. The gold point estimate may reflect trend persistence or defensive episodes, but weak downward recall means it is not balanced two-direction forecasting. The interval is conditional on one continuous-futures feed and is not multiplicity-adjusted; the frozen model should therefore be treated as a prospective hypothesis rather than a proven 55% trading edge.

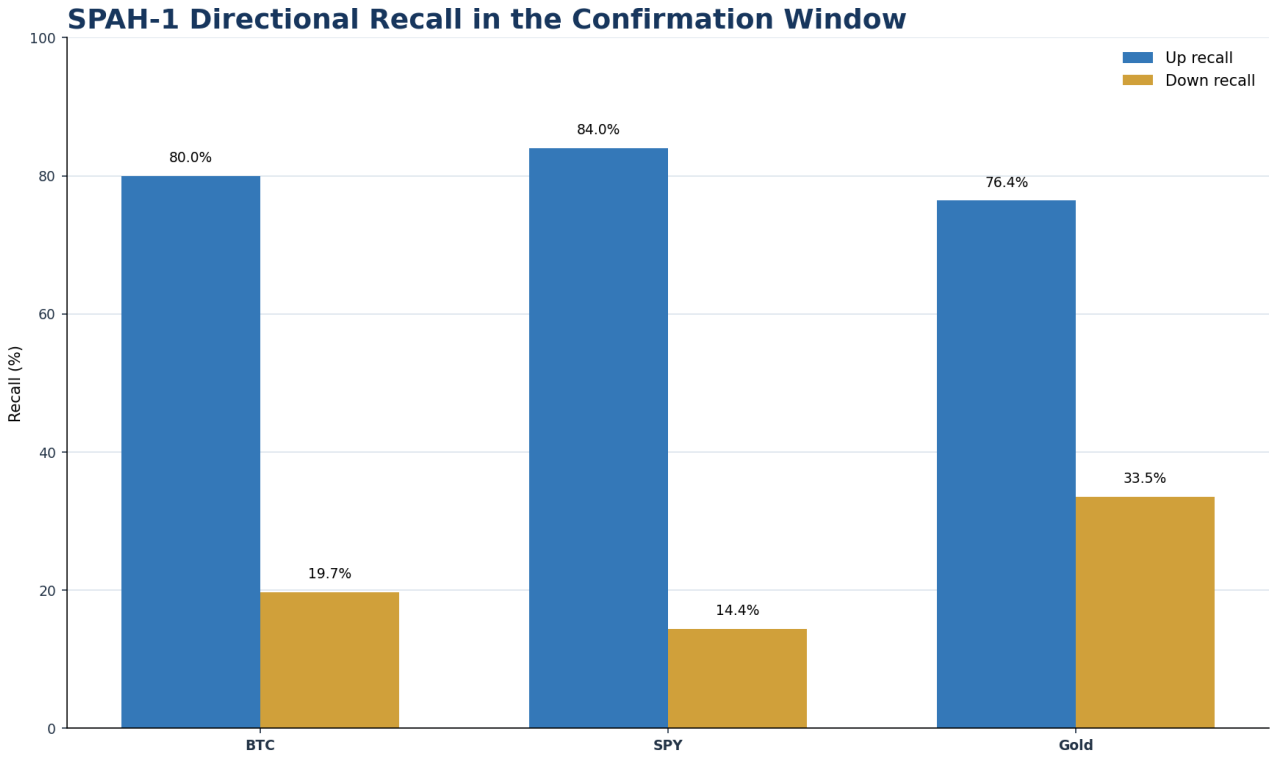

Figure 6. Upward and downward recall of SPAH-1 in the confirmation window.
Source: Authors' calculations.

The candidate architectures and their frozen decisions are detailed in Table 10. Increased complexity does not guarantee better confirmation: the hybrid confirms at 51.3%, the asset-specific model at 51.7%, pooled full at 50.6%, pooled regime at 50.9%, and pooled core at 49.8%. The hybrid was selected before confirmation because it led validation at 51.9%; the higher ex post confirmation of another architecture cannot be used to replace that frozen choice. The full coefficient record is detailed in Table 17.

Table 10. Comparison of predictive architectures; Hybrid is the confirmatory decision.

Architecture	Ridge	Cutoff	Validation	Confirmation	BTC	SPY	Gold
Hybrid Pooled/Asset	100.0	51.0%	51.9%	51.3%	49.8%	49.2%	55.0%
Pooled Full	10.0	52.0%	51.5%	50.6%	49.4%	49.5%	52.9%
Pooled Core	1.0	51.0%	51.5%	49.8%	50.7%	48.8%	50.0%
Asset-Specific Full	1.0	48.0%	51.5%	51.7%	50.7%	50.4%	53.9%
Pooled Regime	1.0	50.0%	51.3%	50.9%	50.0%	50.1%	52.8%

Source: Authors' calculations from the study dataset.

From a modeling perspective, SPAH-1 is preferable to an opaque optimized rule because its training boundary, transformations, coefficients, and cutoff are documented. Nevertheless, transparent methodology does not transform a weak result into a strong one. The main contribution is the disciplined estimate of attainable performance under the tested inputs. A macro balanced accuracy near 51% suggests, at most, a small conditional signal that could easily be offset by implementation frictions or structural change. Claims of precision should therefore focus on the confidence intervals and confirmation protocol rather than the third decimal place of a point estimate.

3.6 Evaluation of the 80% objective

Full-coverage performance by horizon is visualized in Figure 7 and detailed in Table 11. Confirmation accuracy ranges from 47.2% to 57.0%, while BAcc ranges from 49.8% to 53.7%, all far below 80%. At five bars, 57.0% accuracy is accompanied by 99.8% UP recall and 0.0% DOWN recall, showing that the result is almost entirely a majority-class effect. The always-UP benchmark in Table 11 exposes the same imbalance.

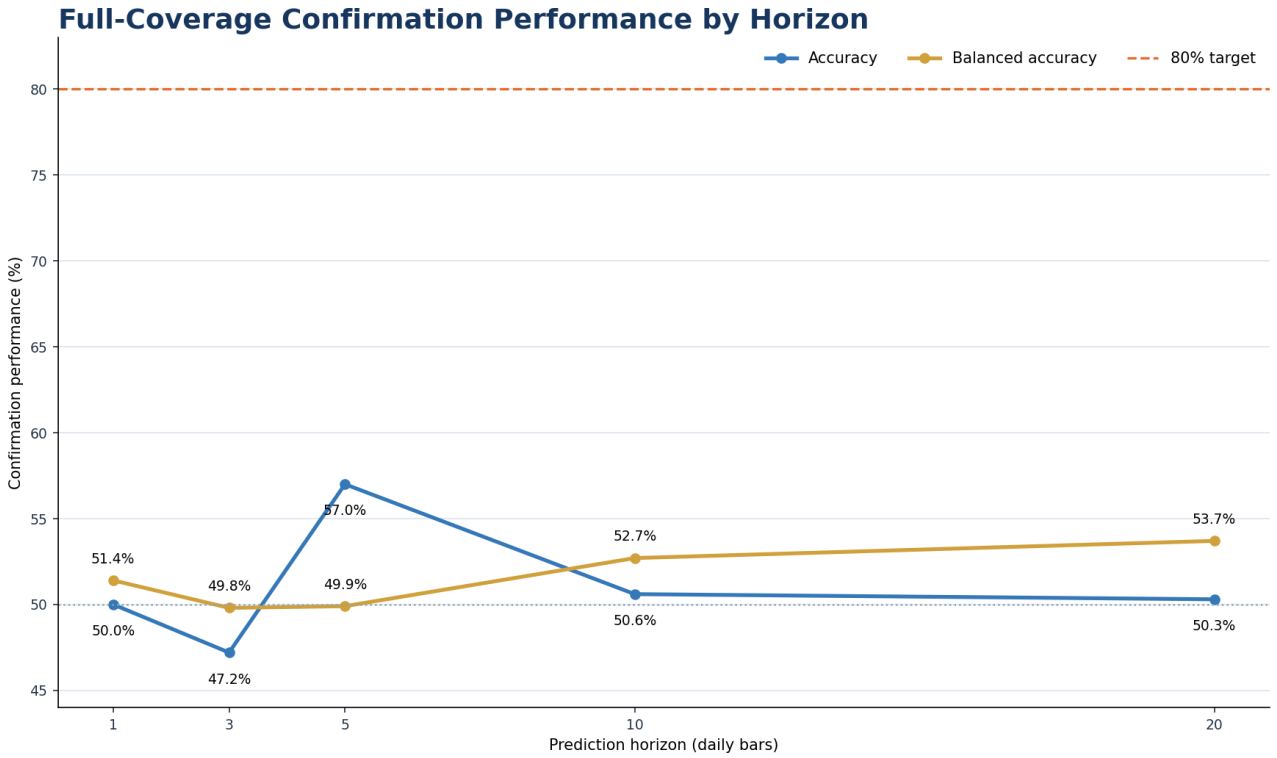

Figure 7. Confirmation accuracy and balanced accuracy of full-coverage models by horizon.
Source: Authors' calculations.

Table 11. Full-coverage model performance by prediction horizon.

Horizon	Model	Accuracy	BAcc	Recall UP	Recall DOWN	Always UP
1 bar	Pooled Regime	50.0%	51.4%	29.5%	73.2%	53.4%
3 bar	Asset-Specific Full	47.2%	49.8%	21.4%	78.2%	56.0%
5 bar	Pooled Regime	57.0%	49.9%	99.8%	0.0%	57.1%
10 bar	Hybrid Pooled/Asset	50.6%	52.7%	33.3%	72.2%	59.2%
20 bar	Pooled Regime	50.3%	53.7%	37.0%	70.4%	61.8%

Source: Authors' calculations from the study dataset.

Selective candidates are detailed in Table 12. The most striking exploratory sensitivity result is a twenty-bar SPY rule with 77.1% accuracy on N=35 non-overlapping executed events: 32 UP and 3 DOWN. Its BAcc is only 59.2%, and its Wilson accuracy interval is 61.0%-87.9%. Table 12 distinguishes this post-selection result from frozen confirmation candidates, including a separate SPY rule that achieves only 55.6% accuracy on N=9. The evidence therefore does not support an 80% balanced two-direction claim.

Table 12. Selective candidates and the exploratory SPY sensitivity result; non-overlapping observations and Wilson 95% intervals.

Asset	Horizon	P(UP) rule	Role	Accuracy/N	BAcc	UP/DOWN	95% CI Acc
BTC	3 bar	DOWN <= 0.43; UP >= 0.65	Frozen confirmation	71.4% / 14	71.4%	11/3	45.4%–88.3%
S&P 500	20 bar	DOWN <= 0.45; UP >= 0.73	Frozen confirmation	55.6% / 9	58.3%	4/5	26.7%–81.1%
S&P 500	20 bar	DOWN <= 0.45; UP >= 0.63	Exploratory sensitivity	77.1% / 35	59.2%	32/3	61.0%–87.9%
Gold	5 bar	DOWN <= 0.39; UP >= 0.70	Frozen confirmation	50.0% / 2	50.0%	2/0	9.5%–90.5%

Source: Authors' calculations from the study dataset.

The addendum therefore falsifies rather than supports the proposed 80% daily target within the tested design. A more defensible research objective is forward BAcc around 60%–65% with a predeclared minimum sample, useful coverage, acceptable turnover, and representation of both directions. Even that target would be ambitious for liquid daily markets. Selective systems should be evaluated as three-state decision processes—UP, DOWN, and NEUTRAL—rather than reporting accuracy only after removing abstentions. Coverage, calibration, selective risk, and opportunity cost belong in the same performance statement as correctness.

4. Discussion

4.1 Practical implications

For practical chart use, the results support a hierarchy rather than a single mechanical trigger. Slow trend measures such as EMA200 and Ichimoku can describe regime. Volume Profile, VWMA, OBV, and MFI can provide confirmation about participation or price location. DSS, Stochastic RSI, and related oscillators may help time an action within a regime, but their crossings should not be treated as standalone next-day forecasts. Risk sizing, stop logic, liquidity, and the investment horizon remain separate decisions. This interpretation is more modest than an automatic buy/sell system, but it is consistent with the evidence.

Composite technical scores can organize heterogeneous signals, yet a combined recommendation inherits the assumptions, correlations, and lags of its components[1–5]. Agreement among correlated moving-average rules is not independent confirmation. Volume measures can also disagree across feeds, especially for decentralized crypto markets and continuous futures. A robust workflow should record the symbol, venue, session, adjustment method, parameter set, and decision timestamp whenever a signal is evaluated.

Researchers extending this work should pre-register a limited family of hypotheses, freeze code before the confirmation date, and retain every tested specification. Alternative vendors and tradable instruments should be used to assess feed dependence. Forecasts should be generated in a live paper environment before capital is exposed. Statistical reporting should include class-specific recall, balanced accuracy, coverage, bootstrap uncertainty, and a correction or explicit accounting for multiple searches. Economic evaluation should incorporate spread, slippage, financing, shorting constraints, taxes, and the distribution of returns conditional on correct and incorrect signals.

4.2 Limitations

Several limitations remain. First, the data feeds are proxies for the exact prices available on other exchanges, venues, and continuous-contract constructions and were not synchronized to one global close. SPY adjustment treatment and the GC=F continuous-contract roll require verification against an archived vendor snapshot. Second, the set of indicators is broad but not exhaustive, and simple directional translations may omit discretionary chart context. Third, the historical period contains a limited number of major regimes, especially for Bitcoin. Fourth, stationary-bootstrap intervals depend on the selected block design and cannot represent future structural breaks. Fifth, no familywise multiplicity correction was applied, so all best-rule rankings and the gold finding remain descriptive. Sixth, the economic simulation lacks a complete executable trade ledger and is stylized. Seventh, the exploratory 80% search examined multiple thresholds and horizons and lacks prospective calibration evidence. Finally, the study evaluates price direction, not return magnitude, utility, or portfolio-level allocation.

Despite these limitations, the design provides a useful falsification exercise. If common indicators possessed a large and stable next-day advantage, the effect should have appeared across assets, survived a frozen confirmation period, and remained visible in class-balanced metrics. It did not. The narrow distribution around 50%, instability of selected parameters, and failure of after-cost strategies align with modern warnings about backtest selection[19–22]. The positive SPAH-1 gold result should be carried forward as a pre-registered hypothesis with a new data snapshot and multiplicity-aware test, not generalized retrospectively.

4.3 Financial education and deployment

The results also have implications for financial education and consumer protection. Screenshots of successful chart signals are easy to select after an event, whereas unsuccessful or ambiguous signals are rarely displayed with equal prominence. A learner may consequently confuse narrative coherence with forecast calibration. Requiring an explicit timestamp, a frozen rule, a predetermined evaluation horizon, and a

complete sequence of predictions would make performance claims substantially more informative. Platform users should also distinguish an indicator's computational correctness from the economic claim attached to it: an RSI or MACD value can be calculated perfectly while the associated prediction remains no better than chance.

A useful deployment protocol would begin with a paper-trading registry. Before each market close, the system would store the input-data hash, computed feature vector, predicted probability, direction or neutral state, and intended execution rule. Outcomes would be scored automatically after the relevant horizon, without deleting signals. Monitoring would use rolling balanced accuracy, calibration, coverage, turnover, and performance by direction and regime. A change to any formula, threshold, symbol, vendor, or cost assumption would create a new model version and a new confirmation period. Such governance cannot guarantee profitability, but it prevents the most common retrospective reinterpretations and makes a small edge easier to distinguish from noise.

5. Conclusion

This study tested whether daily technical indicators predicted the next close-to-close direction of Bitcoin, SPY, and gold from 2015 to 14 July 2026 under a chronological walk-forward design. Signals at close t were evaluated only against the next available close. Individual indicators and predetermined combinations produced BAcc values concentrated near 50%: Ichimoku led the default indicators at 50.9% macro BAcc, and Volume Confirmed led the combinations at 50.8%. The latter underperformed buy-and-hold after costs for all three assets. The principal questions, answers, and supporting values are summarized in Table 13.

Table 13. Summary of principal research questions and evidence.

Question	Answer	Evidence
Does the study use daily observations?	Yes	The signal at close t is matched with the direction of close t+1 relative to close t
Selected predictive model	SPAH-1 Hybrid	Validation 51.9%; macro confirmation 51.3%
SPAH-1 for gold	Promising, not conclusive	55.0%; CI 52.0%–58.2%; down recall remains 33.5%
SPAH-1 for BTC/SPY	Unsuccessful	49.8% BTC and 49.2% SPY in the confirmation window
Best single indicator across assets	Ichimoku 9/26/52	50.9% macro balanced accuracy; small edge, 87.3% coverage
Best combination	Volume Confirmed	50.8% macro balanced accuracy; +0.8 pp versus chance
Best for BTC	Volume Profile POC 60/24	51.1%; CI 49.3%–52.7%
Best for SPY	Bollinger %B 20/2	51.3%; CI 49.3%–53.1%
Best for gold	Ichimoku 9/26/52	51.3%; CI 49.2%–53.5%
Weakest	DSS Bressert 10/3/3	49.3% macro; 48.4% for BTC
Statistical conclusion	No robust edge established	All asset-level winners have confidence intervals that cross 50%

Source: Authors' calculations from the study dataset.

The regularized SPAH-1 hybrid reached 51.9% BAcc in validation and 51.3% in confirmation. Its performance was heterogeneous: Bitcoin was 49.8%, SPY 49.2%, and gold 55.0%, with a strong upward prediction bias. The exploratory 77.1% SPY sensitivity result used only N=35 non-overlapping executed events (32 UP and 3 DOWN) and had BAcc of 59.2%; it was neither balanced nor confirmatory. Longer-horizon and selective experiments therefore failed to justify an 80% accuracy claim.

Technical indicators are more defensible as tools for describing regime, confirming participation, and structuring risk decisions than as universal one-day forecasting engines. Future research should prioritize frozen forward tests, both directional classes, useful coverage, adequate sample size, multiplicity control, and economic outcomes after realistic frictions. Under those requirements, a stable 60%-65% BAcc would already be meaningful. Until such evidence accumulates, claims that common daily indicators can reliably predict the next market direction, especially at 80% accuracy, should be rejected.

Declarations

Funding: This research received no external funding.

Competing interests: The authors declare no competing interests.

Ethics approval: Not applicable. The study uses public secondary market data and involves no human participants, interventions, private information, or identifiable records.

Consent to participate: Not applicable.

Consent for publication: Not applicable.

Data availability: The analysis uses public daily market data identified in Table 1. Exact replication requires the same vendor series, symbols, and freeze date.

Code availability: The manuscript records formulas, parameter settings, bootstrap seed, and frozen coefficients sufficient for independent reimplementation.

Author contributions: Ifah Shandy: conceptualization, methodology, software, formal analysis, investigation, data curation, visualization, writing — original draft. Benyamin Wongso: methodology, validation, formal analysis, investigation, writing — review and editing. All authors have reviewed and approved the final version of the manuscript.

Declaration of generative AI and AI-assisted technologies: The authors declare that no generative artificial intelligence or AI-assisted technology was used in the conception, analysis, or writing of this manuscript. The authors reviewed the methods, data, results, interpretations, and text, and accept responsibility for the publication.

References

1. Mostafavi SM, Hooman AR. Key technical indicators for stock market prediction. Machine Learning with Applications. 2025;20:100631. doi:10.1016/j.mlwa.2025.100631.
2. Peng Y, Albuquerque PHM, Kimura H, et al. Feature selection and deep neural networks for stock price direction forecasting using technical analysis indicators. Machine Learning with Applications. 2021;5:100060. doi:10.1016/j.mlwa.2021.100060.
3. Ma C, Yan S. Deep learning in the Chinese stock market: The role of technical indicators. Finance Research Letters. 2022;49:103025. doi:10.1016/j.frl.2022.103025.
4. Sermpinis G, Hassanniakalager A, Stasinakis C, et al. Technical analysis profitability and persistence: A discrete false discovery approach on MSCI indices. Journal of International Financial Markets, Institutions and Money. 2021;73:101353. doi:10.1016/j.intfin.2021.101353.
5. Ivanova Y, Neely CJ, Weller P, et al. Can risk explain the profitability of technical trading in currency markets? Journal of International Money and Finance. 2021;110:102285. doi:10.1016/j.jimonfin.2020.102285.
6. Máté D, Raza H, Ahmad I, et al. Next step for bitcoin: Confluence of technical indicators and machine learning. Journal of International Studies. 2024;17(3):68-94. doi:10.14254/2071-8330.2023/17-3/4.
7. Omole O, Enke D. Using machine and deep learning models, on-chain data, and technical analysis for predicting bitcoin price direction and magnitude. Engineering Applications of Artificial Intelligence. 2025;154:111086. doi:10.1016/j.engappai.2025.111086.
8. Basher SA, Sadorsky P. Forecasting Bitcoin price direction with random forests: How important are interest rates, inflation, and market volatility? Machine Learning with Applications. 2022;9:100355. doi:10.1016/j.mlwa.2022.100355.
9. Dubey R, Enke D. Bitcoin price direction prediction using on-chain data and feature selection. Machine Learning with Applications. 2025;20:100674. doi:10.1016/j.mlwa.2025.100674.
10. Cortese FP, Kolm PN, Lindström E. What drives cryptocurrency returns? A sparse statistical jump model approach. Digital Finance. 2023;5(3-4):483-518. doi:10.1007/s42521-023-00085-x.
11. Liu Y, Tsyvinski A. Risks and returns of cryptocurrency. Review of Financial Studies. 2021;34(6):2689-2727. doi:10.1093/rfs/hhaa113.
12. Sebastião H, Godinho P. Forecasting and trading cryptocurrencies with machine learning under changing market conditions. Financial Innovation. 2021;7(1):3. doi:10.1186/s40854-020-00217-x.
13. Akyildirim E, Goncu A, Sensoy A. Prediction of cryptocurrency returns using machine learning. Annals of Operations Research. 2021;297(1-2):3-36. doi:10.1007/s10479-020-03575-y.
14. Catania L, Grassi S. Forecasting cryptocurrency volatility. International Journal of Forecasting. 2022;38(3):878-894. doi:10.1016/j.ijforecast.2021.06.005.
15. Gurrib I, Kamalov F. Predicting bitcoin price movements using sentiment analysis: A machine learning approach. Studies in Economics and Finance. 2022;39(3):347-364. doi:10.1108/SEF-07-2021-0293.
16. Cohen G, Aiche A. Forecasting gold price using machine learning methodologies. Chaos, Solitons & Fractals. 2023;175:114079. doi:10.1016/j.chaos.2023.114079.
17. Yang M, Wang R, Zeng Z, et al. Improved prediction of global gold prices: An innovative Hurst-reconfiguration-based machine learning approach. Resources Policy. 2024;88:104430. doi:10.1016/j.resourpol.2023.104430.
18. Azevedo V, Hoegner C. Enhancing stock market anomalies with machine learning. Review of Quantitative Finance and Accounting. 2023;60(1):195-230. doi:10.1007/s11156-022-01099-z.
19. Arian H, Norouzi Mobarekeh D, Seco L. Backtest overfitting in the machine learning era: A comparison of out-of-sample testing methods in a synthetic controlled environment. Knowledge-Based Systems. 2024;305:112477. doi:10.1016/j.knosys.2024.112477.
20. Peng Y, de Moraes Souza JG. Chaos, overfitting and equilibrium: To what extent can machine learning beat the financial market? International Review of Financial Analysis. 2024;95:103474. doi:10.1016/j.irfa.2024.103474.
21. Witzany J. A Bayesian approach to measurement of backtest overfitting. Risks. 2021;9(1):18. doi:10.3390/risks9010018.
22. Ahn H, Jang B. An accurate cryptocurrency price forecasting using reverse walk-forward validation. Journal of Internet Computing and Services. 2022;23(4):45-55. doi:10.7472/jksii.2022.23.4.45.
23. Brodersen KH, Ong CS, Stephan KE, et al. The balanced accuracy and its posterior distribution. In: 2010 20th International Conference on Pattern Recognition. IEEE; 2010. p. 3121-3124. doi:10.1109/ICPR.2010.764.
24. Politis DN, Romano JP. The stationary bootstrap. Journal of the American Statistical Association. 1994;89(428):1303-1313. doi:10.1080/01621459.1994.10476870.