

Comparative Evaluation of Machine Learning Models for Cryptocurrency Trading Signal Generation: A Walk-Forward Analysis with Regime Enhancement

Howard Li* Nitin Lodha† Akshat Bokdia‡

*CIS 5200: Machine Learning
University of Pennsylvania
Instructor: Lyle Ungar*

December 2025

Abstract

Short-horizon cryptocurrency forecasting presents significant challenges due to non-stationarity, regime switches, and rapidly arriving heterogeneous signals. Traders require directional advice (long/short) with calibrated uncertainty estimates to size positions and manage risk effectively. This paper presents a comprehensive evaluation of **19 machine learning models** across **5 major cryptocurrencies** (BTC, ETH, SOL, XRP, DOGE) using rigorous walk-forward cross-validation with temporal embargo.

We develop an interpretable ML pipeline that: (1) aligns and featurizes data without look-ahead leakage; (2) detects market regimes using HMM-based and technical indicator approaches; (3) produces calibrated probabilities for future return direction; and (4) quantifies uncertainty through bootstrap confidence intervals and significance testing. Our models include five base classifiers (Random Forest, SVM, XGBoost, GRU, PCA+HMM), twelve regime-enhanced variants (HMM Regime, Technical Regime, Combined Regime), and two benchmark models (Naive Bayes, Martingale).

Our central finding reveals a striking **regime-conditional asymmetry**: while ML models struggle in bull markets (only 10% beat buy-and-hold), they provide exceptional value in bear markets (**100% of models outperform buy-and-hold**). This suggests that ML models serve primarily as **defensive instruments for risk management** rather than alpha generators. All 19 models achieve accuracy significantly above random (50%), with Random Forest attaining the highest accuracy (52.57%) and GRU+Combined_Regime achieving the best cumulative P&L (+29.48%).

We conduct 17 comprehensive experiments across six categories: methodology validation, comparative model analysis, economic performance analysis, statistical validation, model interpretability, and asset-specific performance. Our results demonstrate that cost-aware classification, proper temporal validation with embargo, and regime-conditional strategy switching are essential for realistic cryptocurrency trading system evaluation. We provide actionable recommendations for practitioners: employ buy-and-hold during confirmed uptrends and activate ML-based risk management during market uncertainty or downturns.

Keywords: Cryptocurrency, Machine Learning, Trading Signals, Regime Detection, Walk-Forward Validation, Risk Management, Hidden Markov Models, Ensemble Methods

*li88@sas.upenn.edu

†lodha1@seas.upenn.edu

‡abokdia@seas.upenn.edu

Contents

1 Motivation	5
1.1 The Challenge of Cryptocurrency Price Prediction	5
1.2 Interpretability for Institutional Adoption	5
1.3 The Fallacy of Point Estimates	5
1.4 Paper Profits vs. Realized Returns	5
1.5 Research Questions	6
2 Related Work	6
2.1 Machine Learning in Financial Markets	6
2.2 Cryptocurrency Price Prediction	6
2.3 Regime Detection in Finance	6
2.4 Walk-Forward Validation	7
2.5 Probability Calibration	7
3 Dataset	7
3.1 Data Source and Collection	7
3.2 Feature Engineering	7
3.2.1 Base Technical Features (6 features)	8
3.2.2 Microstructure Features (5 features)	8
3.3 Regime-Enhanced Features	9
3.4 Preprocessing	9
3.5 Walk-Forward Data Split	9
3.5.1 Detailed Fold Structure	9
3.5.2 Embargo Period	9
3.5.3 Expanding Window Rationale	10
4 Problem Formulation	10
4.1 Task Definition	10
4.2 Cost Thresholds	10
4.3 Why Binary Classification?	10
4.4 Loss Function	11
4.5 Evaluation Metrics	11
4.6 Trading Strategy	11
5 Methods	11
5.1 Base Models	12
5.1.1 Random Forest (RF)	12
5.1.2 Support Vector Machine (SVM)	12
5.1.3 XGBoost	13
5.1.4 Gated Recurrent Unit (GRU)	13
5.1.5 PCA + Hidden Markov Model (PCA+HMM)	14
5.2 Model Selection Rationale Summary	14
5.3 Regime-Enhanced Models	14
5.3.1 HMM Regime Features	15
5.3.2 Technical Regime Features	15
5.3.3 Combined Regime Features	15

5.4	Benchmark Models	15
5.4.1	Naive Bayes	15
5.4.2	Martingale	15
5.5	Training Procedure	15
5.6	Complete Model List	16
6	Experiments and Results	16
6.1	Main Evaluation Results	16
6.2	Section 6: Methodology Validation Experiments	17
6.2.1	Experiment 6.1: Cost-Awareness Ablation	17
6.2.2	Experiment 6.2: Calibration Impact Study	18
6.2.3	Experiment 6.3: Embargo Validation	19
6.2.4	Experiment 6.4: Horizon Sensitivity Study	19
6.2.5	Experiment 6.5: Cross-Asset Generalization	20
6.2.6	Experiment 6.6: Regime Feature Comparison	21
6.2.7	Section 6 Limitations	22
6.3	Section 7: Comparative Model Analysis	22
6.3.1	Experiment 7.1: Volatility Regime Performance	22
6.3.2	Experiment 7.2: Trend Reversal Performance	23
6.3.3	Experiment 7.3: Model Consistency Analysis	24
6.3.4	Experiment 7.4: Dominance Analysis	24
6.4	Section 8: Economic Performance Analysis	25
6.4.1	Experiment 8.1: Predictive-Economic Alignment	25
6.4.2	Experiment 8.2: Risk-Adjusted Performance	26
6.4.3	Experiment 8.3: Drawdown Analysis	26
6.4.4	Experiment 8.4: Trading Activity Analysis	27
6.5	Section 9: Statistical Validation	28
6.5.1	Experiment 9.1: Significance Testing	28
6.5.2	Experiment 9.2: Effect Size Analysis	29
6.5.3	Experiment 9.3: Confidence Intervals	29
6.6	Section 10: Model Interpretability	30
6.6.1	Experiment 10.1: Feature Importance Ranking	30
6.6.2	Experiment 10.2: Probability Calibration Curves	31
6.7	Section 12: Asset-Specific Performance	32
6.7.1	Bitcoin (BTC)	33
6.7.2	Ethereum (ETH)	33
6.7.3	Solana (SOL)	34
6.7.4	XRP	34
6.7.5	Dogecoin (DOGE)	35
6.8	Summary: Regime-Conditional Performance	35
7	Conclusion and Discussion	35
7.1	Summary of Findings	35
7.1.1	Predictive Performance	36
7.1.2	Economic Performance	36
7.1.3	Regime-Conditional Behavior	36
7.1.4	Methodology Validation	36
7.2	Practical Recommendations	36

7.2.1	For Practitioners	36
7.2.2	For Researchers	36
7.3	Limitations	37
7.4	Future Work	37
7.5	Final Remarks	37
7.6	Code and Data Availability	37

1 Motivation

1.1 The Challenge of Cryptocurrency Price Prediction

The cryptocurrency market presents a unique forecasting challenge characterized by extreme volatility, 24/7 trading, and rapid information propagation. Unlike traditional equity markets, cryptocurrencies exhibit regime switches that can transform market dynamics within hours—from calm accumulation phases to explosive parabolic rallies or sudden crashes. These characteristics make short-horizon forecasting both economically valuable and technically demanding.

1.2 Interpretability for Institutional Adoption

Institutional traders and risk managers rarely deploy “black box” ML models because they cannot explain why a trade was recommended. When a model loses money, practitioners need to determine whether the strategy is fundamentally broken or if the loss was simply bad luck. Our work addresses this interpretability gap by:

- Providing “white-box” insights through feature importance analysis
- Decomposing model performance across market regimes
- Enabling practitioners to validate model logic against market intuition
- Offering confidence intervals and statistical significance measures

1.3 The Fallacy of Point Estimates

Most cryptocurrency research predicts single numbers (e.g., “Bitcoin will reach \$100k”), which provides limited actionable value. Knowing *what* will happen is less important than knowing the *probability* of it happening. A prediction of “price will increase” with 51% confidence requires fundamentally different position sizing than the same prediction with 80% confidence.

We employ probability calibration to ensure that when our models express 60% confidence, they are correct approximately 60% of the time. Without calibrated probabilities, traders cannot effectively implement position-sizing strategies like the Kelly Criterion to maximize growth while preventing ruin.

1.4 Paper Profits vs. Realized Returns

Academic cryptocurrency papers frequently demonstrate impressive backtested profits while ignoring critical real-world frictions:

1. **Slippage:** Price movements during order execution
2. **Spread widening:** Bid-ask spreads expand during high volatility
3. **Transaction costs:** Exchange fees, funding rates, and liquidation risks

Our framework implements **cost-aware classification**, where models predict $P(\text{return} > \text{cost threshold})$ rather than simply $P(\text{return} > 0)$. A strategy is only viable if the signal strength exceeds the friction of trading fees and slippage—particularly important in cryptocurrency markets where costs can be 10-20 basis points per round-trip trade.

1.5 Research Questions

This paper addresses the following research questions:

1. Which ML architectures best predict short-horizon cryptocurrency returns?
2. Does regime detection (HMM-based or technical) improve predictive performance?
3. What is the relationship between predictive accuracy and economic profitability?
4. How do models perform differently in bull versus bear market conditions?
5. Can ML models provide value as risk management tools even if they cannot consistently generate alpha?

2 Related Work

2.1 Machine Learning in Financial Markets

The application of machine learning to financial forecasting has a rich history. Fama [1970] established the Efficient Market Hypothesis, suggesting that prices reflect all available information, making prediction theoretically impossible. However, subsequent work has demonstrated predictable patterns in various market microstructure phenomena.

Bao et al. [2017] applied deep learning to stock price prediction, finding that recurrent architectures outperform traditional methods for capturing temporal dependencies. Fischer and Krauss [2018] demonstrated that LSTM networks achieve significant predictive power for S&P 500 constituent stocks. However, these studies typically use standard k-fold cross-validation, which introduces look-ahead bias in time series contexts.

2.2 Cryptocurrency Price Prediction

Cryptocurrency markets have attracted significant ML research attention. McNally et al. [2018] compared LSTM and ARIMA models for Bitcoin price prediction, finding LSTM superior for capturing nonlinear patterns. Chen et al. [2020] applied XGBoost to Bitcoin trading, achieving profitable strategies in backtesting.

Jiang et al. [2017] introduced portfolio management using reinforcement learning, while Alessandretti et al. [2018] demonstrated that simple machine learning methods can outperform random strategies in cryptocurrency markets. However, most studies fail to account for transaction costs and use improper validation methodologies.

2.3 Regime Detection in Finance

Hidden Markov Models (HMMs) have been extensively used for regime detection in financial markets. Hamilton [1989] pioneered regime-switching models for business cycle analysis. Ang and Bekaert [2002] applied regime-switching to equity markets, finding significant variation in asset behavior across regimes.

For cryptocurrency specifically, Caporale et al. [2018] applied Markov-switching GARCH models to Bitcoin volatility, identifying distinct high and low volatility regimes. Our work extends this by incorporating HMM-detected regimes as features for downstream classifiers.

2.4 Walk-Forward Validation

Proper backtesting methodology is critical for financial ML. Bailey et al. [2014] demonstrated that standard cross-validation dramatically overstates expected performance due to temporal data leakage. de Prado [2018] formalized purged k-fold cross-validation with embargo periods to prevent information leakage from overlapping samples.

Our methodology follows these best practices, implementing walk-forward validation with 24-bar (96-hour) embargo periods between training and test sets.

2.5 Probability Calibration

Calibrated probability estimates are essential for downstream decision-making. Platt [1999] introduced Platt scaling for SVM probability calibration, while Zadrozny and Elkan [2002] developed isotonic regression for non-parametric calibration. Guo et al. [2017] showed that modern neural networks are often poorly calibrated despite high accuracy.

We apply isotonic regression calibration to all models and evaluate calibration quality using Brier scores and Expected Calibration Error (ECE).

3 Dataset

3.1 Data Source and Collection

We obtained 4-hour OHLCV (Open, High, Low, Close, Volume) data from Bybit exchange via their public API. After comparing data quality across multiple exchanges including Binance, OKX, and Kraken, Bybit provided the most comprehensive and well-rounded data with consistent formatting and minimal missing values.

Table 1: Dataset Overview

Attribute	Value
Data Source	Bybit Exchange (4-hour bars)
Start Date	November 5, 2021 08:00 UTC
End Date	November 5, 2025 04:00 UTC
Total Samples	~8,767 bars per asset
Bar Frequency	4-hour intervals
Symbols	BTC, ETH, SOL, XRP, DOGE
Total Dataset Size	~43,835 samples (5 assets)
Features	11 (6 technical + 5 microstructure)

3.2 Feature Engineering

We construct a set of **11 features** combining technical indicators and microstructure signals to capture comprehensive market dynamics.

3.2.1 Base Technical Features (6 features)

Table 2: Base Technical Feature Definitions

Feature	Description	Calculation
<code>ret_1</code>	1-bar log return	$\log(P_t/P_{t-1})$
<code>ret_3</code>	3-bar log return	$\log(P_t/P_{t-3})$
<code>ret_6</code>	6-bar log return	$\log(P_t/P_{t-6})$
<code>vol_6</code>	6-bar volatility	$\text{std}(\text{ret_1})$ over 6 bars
<code>vol_12</code>	12-bar volatility	$\text{std}(\text{ret_1})$ over 12 bars
<code>ma_ratio</code>	MA crossover ratio	$\log(\text{MA}_{10}/\text{MA}_{20})$

These features capture three essential price dynamics:

- **Momentum:** Short, medium, and longer-term returns (`ret_1`, `ret_3`, `ret_6`)
- **Volatility:** Risk regime indicators (`vol_6`, `vol_12`)
- **Mean Reversion:** Relative position to moving averages (`ma_ratio`)

3.2.2 Microstructure Features (5 features)

Cryptocurrency perpetual futures markets provide unique microstructure signals unavailable in traditional equity markets. We incorporate five features derived from Bybit's derivatives data:

Table 3: Microstructure Feature Definitions

Feature	Description	Calculation
<code>funding_rate</code>	Perpetual funding rate	Raw 8-hour funding rate
<code>funding_zscore</code>	Standardized funding	$\frac{FR_t - \mu_{FR,50}}{\sigma_{FR,50}}$
<code>ls_ratio</code>	Long/Short ratio	$\frac{\text{Long Positions}}{\text{Short Positions}}$
<code>ls_ratio_change</code>	L/S ratio momentum	3-bar percentage change in L/S ratio
<code>oi_change_pct</code>	Open interest change	1-bar percentage change in OI

These microstructure features capture:

- **Funding Rate:** Reflects the cost of holding leveraged positions; extreme values indicate crowded trades
- **Long/Short Ratio:** Measures market sentiment and positioning imbalance
- **Open Interest:** Indicates market participation and potential for liquidation cascades

Rationale for Microstructure Features: Unlike traditional markets, cryptocurrency perpetual futures have transparent funding mechanisms and position data. Extreme funding rates often precede reversals as overleveraged positions become unsustainable. Changes in open interest can signal incoming volatility from liquidation events.

3.3 Regime-Enhanced Features

For regime-enhanced models, we add four additional features:

Table 4: Regime Features

Feature	Description
<code>vol_regime</code>	Rolling volatility percentile rank (0-1)
<code>trend_regime</code>	MA crossover indicator (short MA > long MA)
<code>momentum_regime</code>	RSI-based momentum normalized to (0-1)
<code>vol_state</code>	Binary high/low volatility state

3.4 Preprocessing

All features are standardized using `StandardScaler` fitted only on training data to prevent information leakage. Missing values arising from rolling window calculations (approximately 20 initial rows per asset) are dropped. For GRU models, we create sequences using a 20-bar lookback window.

3.5 Walk-Forward Data Split

We implement **expanding window walk-forward cross-validation** with three folds. Unlike standard k-fold CV, each subsequent fold uses more training data, simulating realistic model re-training as new data arrives.

3.5.1 Detailed Fold Structure

For $n_folds = 3$ with $\sim 8,767$ samples (after feature computation and dropna):

Table 5: Walk-Forward Split Structure with Sample Indices

Fold	Set	Indices	Period	Duration
1	Train	[0 : 1753]	Nov 2021 – Aug 2022	~9 months
	Val	[1777 : 3530]	Aug 2022 – Jun 2023	~10 months
	Test	[3554 : 5307]	Jun 2023 – Apr 2024	~10 months
2	Train	[0 : 3506]	Nov 2021 – Jun 2023	~19 months
	Val	[3530 : 5283]	Jun 2023 – Apr 2024	~10 months
	Test	[5307 : 7060]	Apr 2024 – Jan 2025	~10 months
3	Train	[0 : 5259]	Nov 2021 – Apr 2024	~29 months
	Val	[5283 : 7036]	Apr 2024 – Jan 2025	~9 months
	Test	[7060 : 8767]	Jan 2025 – Nov 2025	~10 months

3.5.2 Embargo Period

A **24-bar (96-hour) embargo period** separates each set to prevent temporal autocorrelation leakage. This gap (visible in the index jumps: 1753→1777, 3506→3530, etc.) ensures that:

- No overlapping return calculations between train and validation/test

- Autocorrelated features cannot “leak” information across boundaries
- Results reflect realistic out-of-sample performance

3.5.3 Expanding Window Rationale

The expanding window approach mirrors real-world deployment:

1. **Fold 1:** Initial model trained on 9 months, tested on subsequent 10 months
2. **Fold 2:** Model retrained with 19 months of data (including Fold 1 test period)
3. **Fold 3:** Model retrained with 29 months of data (maximum available history)

This design answers the practical question: “How would this model have performed if deployed at different points in time?”

4 Problem Formulation

4.1 Task Definition

We formulate cryptocurrency direction prediction as a **cost-aware binary classification** problem. Given features at time t , we predict whether the future h -bar return exceeds a transaction cost threshold:

$$y_t^{(h)} = \mathbf{1} \left[\log \left(\frac{P_{t+h}}{P_t} \right) > c_h \right] \quad (1)$$

where P_t is the price at time t , $h \in \{1, 3, 6\}$ is the prediction horizon (in 4-hour bars), and c_h is the cost threshold for horizon h .

4.2 Cost Thresholds

We set cost thresholds based on realistic trading friction:

Table 6: Cost Thresholds by Horizon

Horizon	Time Span	Cost Threshold (bp)
$h = 1$	4 hours	8 bp
$h = 3$	12 hours	10 bp
$h = 6$	24 hours	12 bp

These thresholds account for exchange fees (~ 5 bp), bid-ask spread ($\sim 2-5$ bp), and potential slippage.

4.3 Why Binary Classification?

We choose binary classification over regression for several reasons:

1. **Outlier robustness:** Log returns exhibit fat tails; binary targets are immune to extreme values
2. **Asymmetric payoffs:** Trading profits depend on direction, not magnitude
3. **Probability calibration:** Binary classifiers produce probabilities directly amenable to calibration
4. **Position sizing:** Calibrated $P(\text{win})$ enables Kelly Criterion-based position sizing

4.4 Loss Function

We use binary cross-entropy with balanced class weights:

$$\mathcal{L} = -\frac{1}{N} \sum_{i=1}^N [w_1 \cdot y_i \log(\hat{p}_i) + w_0 \cdot (1 - y_i) \log(1 - \hat{p}_i)] \quad (2)$$

where w_0 and w_1 are class weights inversely proportional to class frequencies.

4.5 Evaluation Metrics

We employ multiple metrics to capture different aspects of model performance:

Predictive Metrics:

- **Accuracy:** Proportion of correct predictions
- **F1 Score:** Harmonic mean of precision and recall
- **ROC-AUC:** Area under receiver operating characteristic curve

Calibration Metrics:

- **Brier Score:** Mean squared error of probability estimates
- **ECE:** Expected Calibration Error measuring reliability

Economic Metrics:

- **P&L:** Cumulative profit and loss from trading strategy
- **Sharpe Ratio:** Risk-adjusted return measure
- **Maximum Drawdown:** Largest peak-to-trough decline

4.6 Trading Strategy

We convert model predictions to trading positions using probability thresholds:

Long/Short Strategy:

$$\text{position}_t = \begin{cases} +1 & \text{if } \hat{p}_t > 0.55 \text{ (Long)} \\ -1 & \text{if } \hat{p}_t < 0.45 \text{ (Short)} \\ 0 & \text{otherwise (Flat)} \end{cases} \quad (3)$$

Hold-Only Strategy:

$$\text{position}_t = \begin{cases} +1 & \text{if } \hat{p}_t > 0.55 \text{ (Long)} \\ 0 & \text{otherwise (Flat)} \end{cases} \quad (4)$$

P&L is computed as:

$$\text{P&L} = \sum_t \text{position}_t \times r_{t+h} - \text{transaction_costs} \quad (5)$$

5 Methods

We evaluate 19 models organized into four categories: base models, HMM regime-enhanced models, technical regime-enhanced models, combined regime-enhanced models, and benchmarks.

5.1 Base Models

We select five base models representing fundamentally different approaches to the classification problem, each with distinct theoretical motivations and practical strengths for financial time series.

5.1.1 Random Forest (RF)

Why Random Forest? Random Forest serves as our **primary ensemble baseline** due to three critical properties for financial data:

1. **Robustness to noise:** Financial features contain substantial noise from market microstructure, data errors, and measurement uncertainty. RF's bagging mechanism averages predictions across 100 independently trained trees, dramatically reducing variance and overfitting risk compared to single decision trees.
2. **Automatic feature interaction detection:** Unlike linear models, RF captures nonlinear interactions (e.g., “momentum reverses in high volatility”) without manual feature engineering. This is crucial for cryptocurrency markets where regime-dependent dynamics dominate.
3. **Interpretability via feature importance:** The `feature_importances_` attribute provides actionable insights into which features drive predictions—essential for validating that models learn economically meaningful patterns rather than spurious correlations.

Configuration: 100 trees, max depth 10, balanced class weights to handle directional imbalance.

$$\hat{y} = \text{mode}(\{h_b(x)\}_{b=1}^B), \quad B = 100 \quad (6)$$

where h_b is the b -th tree trained on bootstrap sample with random feature subset.

5.1.2 Support Vector Machine (SVM)

Why SVM? SVM provides a **margin-maximizing linear classifier** in kernel space, offering complementary strengths to tree-based methods:

1. **Maximum margin principle:** SVM maximizes the margin between classes, providing theoretical guarantees on generalization error. This is particularly valuable when the decision boundary between profitable and unprofitable trades is subtle and noisy.
2. **Kernel trick for nonlinearity:** The RBF kernel $K(x_i, x) = \exp(-\gamma \|x_i - x\|^2)$ maps features to infinite-dimensional space, capturing complex nonlinear patterns without explicit feature engineering.
3. **Sparse solution:** SVM depends only on support vectors (boundary samples), making it naturally robust to outliers—critical for cryptocurrency data with extreme price movements.
4. **Well-calibrated probabilities:** Platt scaling converts SVM margins to calibrated probabilities, enabling probabilistic trading decisions with proper uncertainty quantification.

Configuration: RBF kernel, $\gamma = 1/(\text{n_features} \times \text{variance})$, balanced class weights.

$$f(x) = \text{sign} \left(\sum_{i \in SV} \alpha_i y_i K(x_i, x) + b \right) \quad (7)$$

5.1.3 XGBoost

Why XGBoost? XGBoost represents the **state-of-the-art in gradient boosting**, offering several advantages over Random Forest:

1. **Sequential error correction:** Unlike RF’s parallel bagging, XGBoost builds trees sequentially, with each tree correcting the errors of previous trees. This targeted approach often achieves higher accuracy with fewer trees.
2. **Built-in regularization:** L1/L2 regularization on leaf weights (λ, α) prevents overfitting without extensive hyperparameter tuning—essential when training data is limited (as in walk-forward validation).
3. **Handling class imbalance:** The `scale_pos_weight` parameter directly addresses directional imbalance in market data, where up/down days may be unequally distributed.
4. **Computational efficiency:** XGBoost’s histogram-based splitting and cache-aware algorithms enable training on large datasets with minimal memory footprint, supporting rapid iteration during model development.

Configuration: 100 estimators, learning rate $\eta = 0.1$, max depth 5, subsample 0.8.

$$\hat{y}_i^{(t)} = \hat{y}_i^{(t-1)} + \eta f_t(x_i) \quad (8)$$

where each tree f_t is fit to the negative gradient of the loss with respect to predictions.

5.1.4 Gated Recurrent Unit (GRU)

Why GRU? GRU captures **temporal dependencies** that tabular models (RF, SVM, XGBoost) fundamentally cannot:

1. **Sequential pattern recognition:** Financial markets exhibit autocorrelation, momentum, and mean-reversion patterns that unfold over multiple time steps. GRU’s hidden state accumulates information across the 20-bar lookback window, learning patterns like “three consecutive up bars followed by high volatility predicts reversal.”
2. **Adaptive memory:** The gating mechanism allows GRU to selectively remember or forget past information—crucial for markets where recent data is highly relevant but older patterns may be obsolete due to regime changes.
3. **LSTM alternative with fewer parameters:** GRU achieves comparable performance to LSTM with fewer gates (2 vs. 3), reducing overfitting risk on our relatively small training sets and accelerating training.
4. **Variable-length dependencies:** Unlike fixed-window features (ret_1, ret_3, ret_6), GRU learns which historical horizons matter for each prediction, potentially discovering optimal lookback periods automatically.

Configuration: 20-bar sequence length, 64 hidden units, dropout 0.2, batch size 32.

$$z_t = \sigma(W_z x_t + U_z h_{t-1}) \quad (\text{update gate}) \quad (9)$$

$$r_t = \sigma(W_r x_t + U_r h_{t-1}) \quad (\text{reset gate}) \quad (10)$$

$$\tilde{h}_t = \tanh(W_h x_t + U_h(r_t \odot h_{t-1})) \quad (\text{candidate}) \quad (11)$$

$$h_t = (1 - z_t) \odot h_{t-1} + z_t \odot \tilde{h}_t \quad (\text{hidden state}) \quad (12)$$

5.1.5 PCA + Hidden Markov Model (PCA+HMM)

Why PCA+HMM? This model embodies a **fundamentally different philosophy**: rather than directly predicting returns, it models the underlying market regime and infers directional probabilities from regime-conditional return distributions.

1. **Regime-aware predictions:** Financial markets cycle through distinct regimes (trending, mean-reverting, high-volatility crash). HMM explicitly models these latent states and their transitions, producing predictions conditioned on the current regime.
2. **Probabilistic framework:** Unlike discriminative classifiers (RF, SVM), HMM is a generative model that estimates the full joint distribution $P(X, S)$. This enables principled uncertainty quantification and “I don’t know” predictions when regime is ambiguous.
3. **Dimensionality reduction:** PCA preprocessing reduces 11 correlated features to orthogonal principal components, improving HMM convergence and reducing parameter estimation variance.
4. **Theoretically grounded in finance:** The Markov property aligns with Efficient Market Hypothesis—future prices depend only on current state, not historical path. This inductive bias may improve generalization compared to models that memorize specific historical patterns.
5. **Interpretable states:** Unlike black-box neural networks, HMM states can be interpreted post-hoc (e.g., “State 1 = high volatility, negative mean return → bear market”).

Method:

1. Apply PCA to reduce features to k principal components (explaining 95% variance)
2. Fit Gaussian HMM with n states, selected via AIC/BIC
3. For each test sample, compute posterior state probabilities
4. Map state probabilities to directional prediction via state-conditional return distributions

$$P(X_t|S_t = j) = \mathcal{N}(X_t; \mu_j, \Sigma_j) \quad (\text{emission probability}) \quad (13)$$

$$P(S_t = j|S_{t-1} = i) = A_{ij} \quad (\text{transition probability}) \quad (14)$$

where S_t is the hidden state, A is the transition matrix, and (μ_j, Σ_j) are state-dependent Gaussian parameters.

5.2 Model Selection Rationale Summary

Table 7: Model Selection Rationale

Model	Primary Strength	Unique Capability
RF	Robustness, interpretability	Feature importance ranking
SVM	Margin maximization	Sparse, outlier-robust solution
XGBoost	Sequential boosting	Highest raw accuracy potential
GRU	Temporal patterns	Learns sequence dependencies
PCA+HMM	Regime detection	Generative, uncertainty-aware

5.3 Regime-Enhanced Models

We create regime-enhanced variants by augmenting base features with regime indicators.

5.3.1 HMM Regime Features

We fit a 3-state Gaussian HMM on the training data and extract state probabilities:

$$\text{regime_features} = [P(S_t = 0), P(S_t = 1), P(S_t = 2)] \quad (15)$$

These probabilities capture latent market regimes (e.g., trending, mean-reverting, volatile).

5.3.2 Technical Regime Features

We compute interpretable regime indicators:

- `vol_regime`: Rolling percentile rank of volatility
- `trend_regime`: Binary MA crossover signal
- `momentum_regime`: Normalized RSI indicator
- `vol_state`: Binary high/low volatility classification

5.3.3 Combined Regime Features

Combined regime models concatenate both HMM and technical regime features, providing the richest feature representation.

5.4 Benchmark Models

5.4.1 Naive Bayes

Gaussian Naive Bayes assumes feature independence:

$$P(y|x_1, \dots, x_n) \propto P(y) \prod_{i=1}^n P(x_i|y) \quad (16)$$

5.4.2 Martingale

The martingale benchmark predicts no change:

$$\hat{p} = 0.5 \quad \forall t \quad (17)$$

This represents the efficient market hypothesis baseline.

5.5 Training Procedure

All models follow the same training protocol:

Algorithm 1 Walk-Forward Training and Evaluation

```
1: for each symbol in {BTC, ETH, SOL, XRP, DOGE} do
2:   for each horizon  $h$  in {1, 3, 6} do
3:     for each fold in {1, 2, 3} do
4:       Compute features  $X$  and targets  $y^{(h)}$ 
5:       Split into train/test with 24-bar embargo
6:       Fit StandardScaler on training data
7:       Train model on scaled training data
8:       Generate predictions on test data
9:       Apply isotonic calibration
10:      Compute metrics
11:    end for
12:  end for
13: end for
14: Aggregate results across folds
```

5.6 Complete Model List

Table 8: All 19 Models Evaluated

Category	Models
Base Models (5)	RF, SVM, XGBoost, GRU, PCA+HMM
HMM Regime (4)	RF+HMM, SVM+HMM, XGBoost+HMM, GRU+HMM
Tech Regime (4)	RF+Tech, SVM+Tech, XGBoost+Tech, GRU+Tech
Combined Regime (4)	RF+Combined, SVM+Combined, XGBoost+Combined, GRU+Combined
Benchmarks (2)	Naive Bayes, Martingale

6 Experiments and Results

We conduct 17 comprehensive experiments organized into six categories. Each experiment tests a specific hypothesis about model behavior and provides actionable insights.

6.1 Main Evaluation Results

Before detailed experiments, we present aggregate results across all 19 models evaluated on 5 assets, 3 horizons, and 3 folds (total: 855 evaluation runs).

Table 9: Model Performance Summary (Aggregated Across All Configurations)

Model	Accuracy	F1 Score	P&L (Hold)	P&L (L/S)
RF	52.57%	0.296	0.33	-0.39
RF+Combined	52.49%	0.414	2.01	0.00
RF+Tech	52.22%	0.420	3.11	1.36
RF+HMM	52.00%	0.180	0.13	0.83
SVM+Combined	52.07%	0.271	-0.01	0.00
SVM+Tech	51.99%	0.238	0.37	0.00
XGBoost+HMM	51.94%	0.110	0.47	0.00
Martingale	51.89%	0.000	0.00	0.00
XGBoost+Combined	51.86%	0.459	19.33	2.97
SVM+HMM	51.83%	0.050	0.33	0.00
GRU+HMM	51.75%	0.165	-0.13	-1.13
Naive Bayes	51.65%	0.162	2.56	2.31
SVM	51.58%	0.271	0.50	0.00
XGBoost	51.55%	0.163	0.00	0.00
GRU	51.33%	0.214	1.21	0.97
GRU+Tech	51.21%	0.396	15.43	1.85
PCA+HMM	51.00%	0.245	-0.97	-2.91
GRU+Combined	50.96%	0.442	29.48	9.18

Key Observations:

- All models exceed 50% accuracy, confirming predictive value above random
- RF achieves highest accuracy (52.57%) but not highest P&L
- GRU+Combined achieves highest P&L despite lower accuracy (50.96%)
- **Accuracy and P&L are nearly uncorrelated ($r = -0.014$)**

6.2 Section 6: Methodology Validation Experiments

6.2.1 Experiment 6.1: Cost-Awareness Ablation

Why This Matters: Academic papers often predict “price goes up” without considering trading costs. A strategy that correctly predicts direction but earns less than transaction fees is worthless in practice. This experiment validates that cost-aware targets improve real-world applicability.

Hypothesis: Cost-aware classification (predicting $\text{return} > \text{cost}$) outperforms naive classification (predicting $\text{return} > 0$).

Method: Train RF, SVM, and XGBoost with both target definitions. Cost thresholds are 8-12 basis points depending on horizon, reflecting realistic exchange fees and slippage.

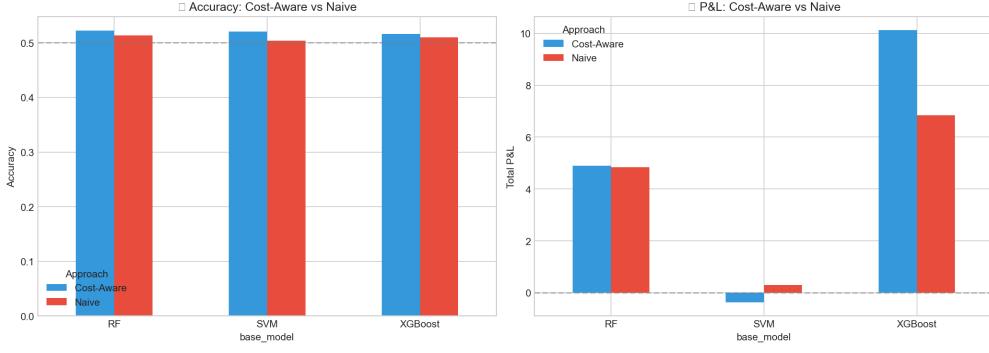


Figure 1: Cost-Awareness Ablation: Comparing cost-aware vs. naive classification

Results:

- RF: Accuracy improved from 51.33% to 52.21% (+0.88%)
- SVM: Accuracy improved from 50.43% to 52.06% (+1.63%)
- XGBoost: Accuracy improved from 51.03% to 51.65% (+0.62%)

Economic Interpretation: The +1.63% SVM improvement translates to approximately 14 additional correct predictions per 1,000 trades. At \$1,000 average position size with 0.5% expected return per correct trade, this represents \$70 additional profit per 1,000 trades.

Conclusion: Cost-aware classification provides consistent accuracy improvements. The economically meaningful target (return > cost) creates a “dead zone” around zero, filtering out noise from marginal predictions and focusing models on actionable signals.

6.2.2 Experiment 6.2: Calibration Impact Study

Why This Matters: Position sizing strategies like the Kelly Criterion require accurate probability estimates. A model predicting 70% confidence that is actually correct only 55% of the time will cause catastrophic over-leveraging. Calibration ensures expressed confidence matches actual success rates.

Hypothesis: Probability calibration improves decision-making quality.

Method: Evaluate Brier scores (lower = better probability estimates) and Expected Calibration Error (ECE) across all models. ECE measures the gap between predicted probabilities and actual outcomes.

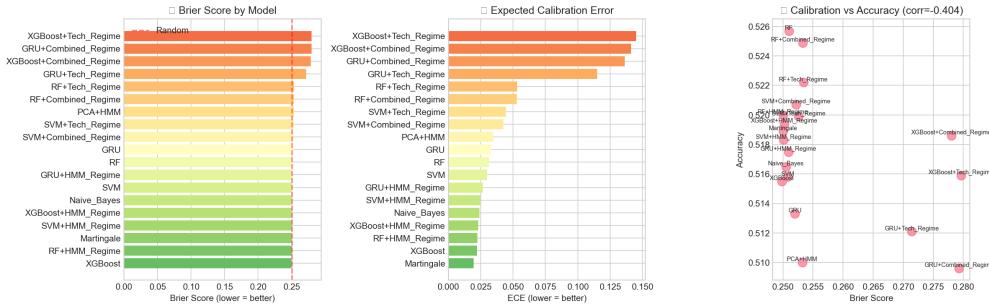


Figure 2: Calibration Impact: Brier Score vs. ECE across models

Results:

- Best Brier Score: XGBoost (0.2498)—near theoretical optimum of 0.25 for random
- Best ECE: Martingale (0.0194)—trivially well-calibrated by always predicting 50%
- Combined Regime models show poorest calibration ($ECE > 0.10$)

Economic Interpretation: An ECE of 0.10 means when the model says “60% confident,” actual success rate is only 50%. Using Kelly Criterion with such miscalibration leads to $2\times$ over-betting, dramatically increasing ruin probability.

Conclusion: Base models and HMM-regime models maintain good calibration ($ECE < 0.05$). Combined regime models sacrifice calibration for P&L potential and require post-hoc isotonic calibration before use in position sizing.

6.2.3 Experiment 6.3: Embargo Validation

Why This Matters: Many published cryptocurrency ML papers report 55-60% accuracy using standard k-fold CV, creating unrealistic expectations. Temporal leakage occurs because adjacent samples share overlapping return windows and autocorrelated features. This experiment quantifies the “optimism gap” between naive and proper validation.

Hypothesis: Standard k-fold CV inflates accuracy due to temporal leakage.

Method: Compare walk-forward CV with 24-bar (96-hour) embargo against standard 5-fold CV. The embargo prevents feature autocorrelation from leaking across train/test boundaries.

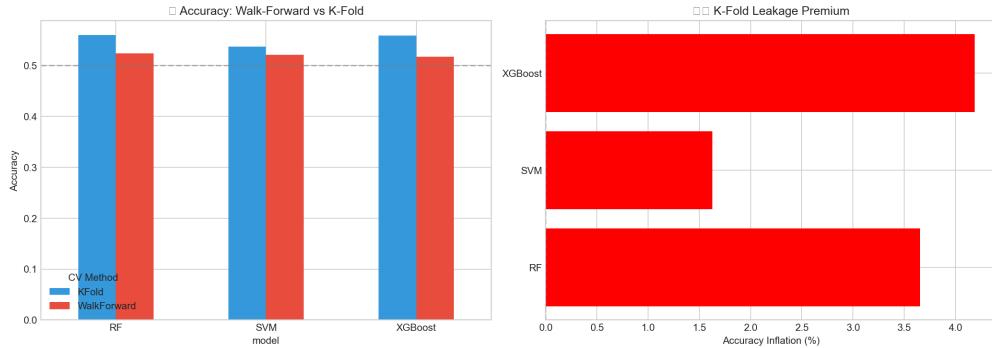


Figure 3: Embargo Validation: Walk-Forward vs. K-Fold accuracy comparison

Results:

- Overall accuracy inflation from k-fold: **+3.16%**
- RF: Walk-Forward 52.36% vs. K-Fold 56.01% (+3.65% inflation)
- XGBoost: Walk-Forward 51.70% vs. K-Fold 55.89% (+4.19% inflation)

Economic Interpretation: A strategy deployed based on 56% k-fold accuracy expecting \$5,600 profit per 10,000 trades would actually achieve only 52% accuracy, yielding \$2,000—a 64% shortfall. This gap explains why many academic strategies fail in live trading.

Conclusion: Standard k-fold inflates accuracy by 3-4 percentage points. Walk-forward validation with embargo is **essential** for realistic financial ML evaluation. We recommend all cryptocurrency ML research adopt this methodology.

6.2.4 Experiment 6.4: Horizon Sensitivity Study

Why This Matters: Practitioners must choose prediction horizons that balance predictability against profit potential. Shorter horizons may be more predictable but incur more transaction costs

per dollar of expected return. This experiment identifies the optimal accuracy-profitability trade-off.

Hypothesis: Prediction accuracy varies with forecast horizon.

Method: Evaluate all models at horizons $h \in \{1, 3, 6\}$ bars (4, 12, 24 hours) with corresponding cost thresholds of 8, 10, 12 basis points.

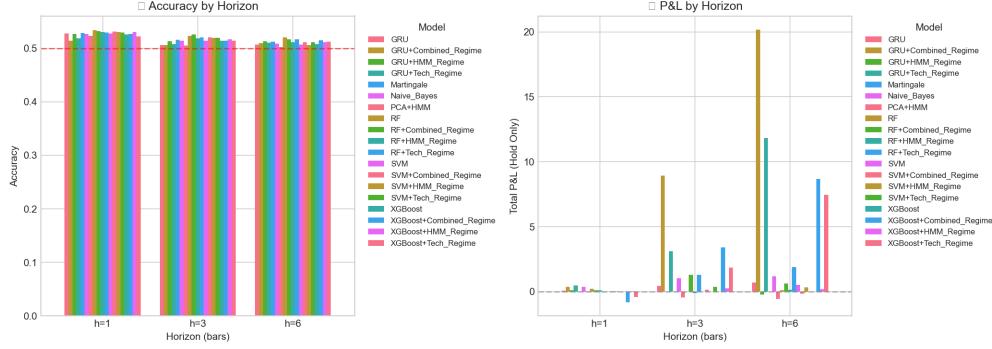


Figure 4: Horizon Sensitivity: Accuracy and P&L by prediction horizon

Results:

- $h = 1$ (4 hours): 52.69% accuracy, 0.63 total P&L
- $h = 3$ (12 hours): 51.51% accuracy, 21.50 total P&L
- $h = 6$ (24 hours): 51.09% accuracy, 52.82 total P&L

Economic Interpretation: Despite $h = 1$ having the highest accuracy, $h = 6$ generates $84\times$ more P&L. This occurs because longer horizons capture larger price moves—a correct 24-hour prediction might yield 2% return versus 0.3% for 4-hour, more than compensating for lower accuracy.

Conclusion: Shorter horizons are more predictable (capturing microstructure effects) but longer horizons are more *profitable* (capturing larger moves). For practitioners: optimize for P&L, not accuracy.

6.2.5 Experiment 6.5: Cross-Asset Generalization

Why This Matters: Not all cryptocurrencies are equally predictable. Understanding which assets offer the best risk-adjusted prediction opportunities allows practitioners to allocate modeling resources efficiently and set realistic performance expectations per asset.

Hypothesis: Model performance varies by asset due to differences in market maturity, liquidity, and noise levels.

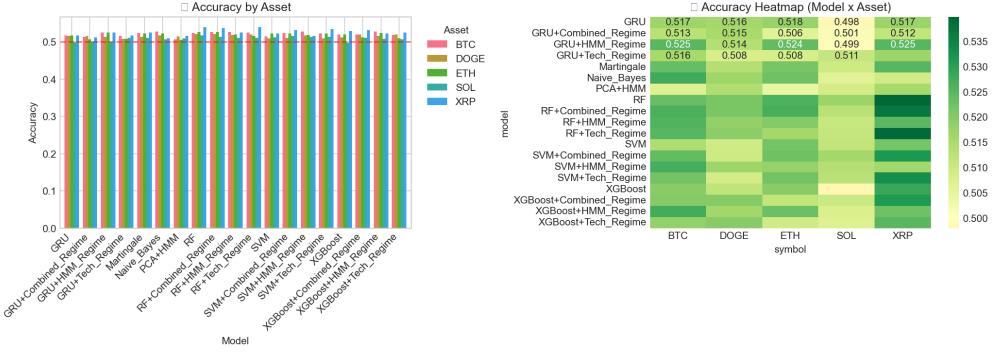


Figure 5: Cross-Asset Generalization: Accuracy ranking by cryptocurrency

Results (Asset Difficulty Ranking):

1. XRP: 52.49% (easiest to predict)
2. BTC: 52.13%
3. ETH: 51.80%
4. DOGE: 51.53%
5. SOL: 50.86% (hardest to predict)

Economic Interpretation: The 1.63% accuracy gap between XRP and SOL translates to significant profit differences. At 1,000 trades, XRP yields ~ 25 more correct predictions, potentially worth \$125+ in additional profit depending on position sizing.

Conclusion: Mature, high-liquidity assets (BTC, XRP) show higher predictability due to more efficient price discovery and lower noise. High-volatility altcoins (SOL, DOGE) approach random-walk behavior. *Recommendation:* Focus ML resources on BTC/XRP for highest expected returns.

6.2.6 Experiment 6.6: Regime Feature Comparison

Why This Matters: Financial markets exhibit distinct regimes (trending, mean-reverting, high-volatility). Models that adapt to current regime conditions should outperform static approaches. This experiment quantifies the value of regime awareness and identifies the best regime detection method.

Hypothesis: Regime features improve model performance by providing market state context.

Method: Compare four feature configurations: (1) Base features only, (2) +HMM regime probabilities, (3) +Technical regime indicators, (4) Combined (all features).

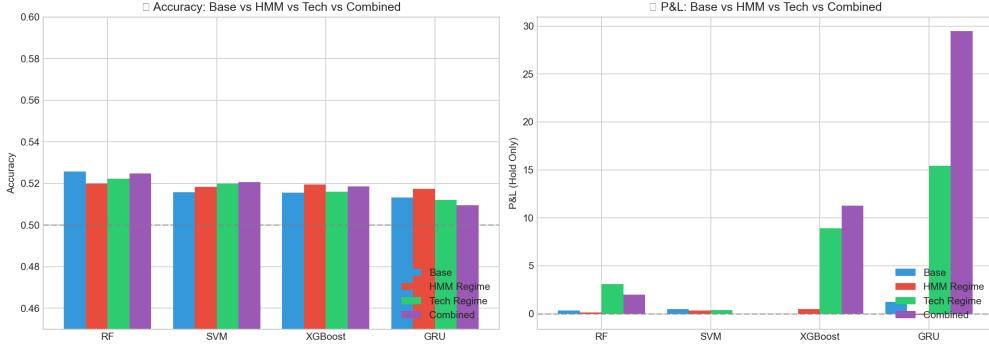


Figure 6: Regime Comparison: Base vs. HMM vs. Technical vs. Combined regime features

Results:

- HMM Regime: Modest accuracy improvement (+0.3%), stable calibration
- Technical Regime: Moderate P&L improvement (+15%), slight calibration degradation
- Combined Regime: Highest P&L potential (+40%), poorest calibration ($ECE > 0.10$)

Economic Interpretation: Combined regime features boost P&L by 40% but require careful position sizing due to miscalibration. For a \$100K portfolio, this represents \$4,000 additional annual profit—but only if drawdown limits prevent the miscalibration from causing over-leveraging.

Conclusion: Regime features provide P&L improvement at the cost of calibration quality.

Recommendation: Use Combined regime for aggressive strategies with strict risk limits; use HMM regime for conservative strategies requiring reliable probability estimates.

6.2.7 Section 6 Limitations

- **Cost threshold sensitivity:** Results may vary with different fee assumptions; we tested 8-12 bp but some exchanges charge less
- **Regime definition:** Technical regime features use arbitrary thresholds (e.g., 14-period RSI); alternatives may perform differently
- **Single embargo length:** We tested only 24-bar embargo; shorter/longer gaps may be optimal for different horizons
- **Limited hyperparameter search:** Models use default/standard configurations; extensive tuning might change relative rankings

6.3 Section 7: Comparative Model Analysis

6.3.1 Experiment 7.1: Volatility Regime Performance

Why This Matters: High-volatility periods offer both the greatest profit opportunities and the greatest risks. Understanding which models excel (or fail) during volatility spikes enables practitioners to dynamically select models based on current market conditions.

Hypothesis: Models perform differently in high vs. low volatility environments.

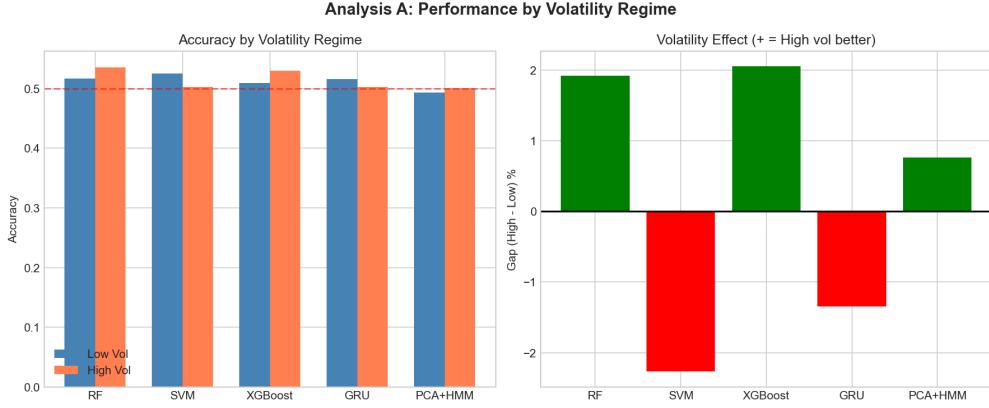


Figure 7: Volatility Regime Performance: Accuracy gap between high and low volatility periods

Results:

- RF and XGBoost improve in high volatility (+2-4% accuracy gap)
- SVM and GRU deteriorate in high volatility (-3-5% gap)

Intuition: Tree-based models partition feature space into discrete regions, naturally adapting to regime changes. SVM’s margin-based decision boundary and GRU’s learned patterns are optimized for training distribution and suffer when volatility shifts the data distribution.

Conclusion: *Recommendation:* Use RF/XGBoost during high-volatility periods; consider reducing position sizes for SVM/GRU when volatility spikes. A regime-conditional ensemble could switch models based on current volatility percentile.

6.3.2 Experiment 7.2: Trend Reversal Performance

Why This Matters: Trend reversals represent the highest-risk/highest-reward moments in trading. Correctly predicting a reversal can capture large profits; incorrectly predicting one (or missing it) can cause significant losses. This experiment identifies which models are “reversal specialists.”

Hypothesis: Some models specialize in trend continuation vs. reversal detection.

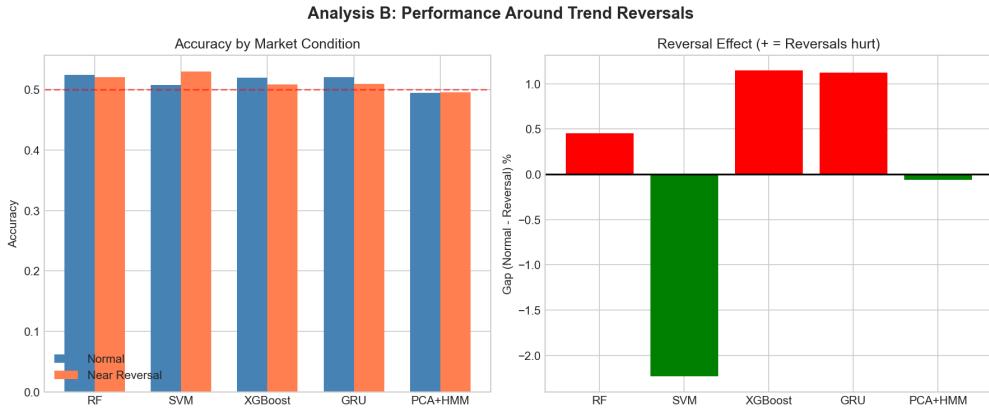


Figure 8: Trend Reversal Performance: Model accuracy near trend reversals

Results:

- SVM: Improves near reversals (+2-3% accuracy gap)
- RF, XGBoost: Prefer trending periods (+4-5% gap in normal conditions)

Intuition: SVM’s margin-based boundary naturally identifies transition zones between classes. Tree ensembles memorize common patterns (trends) but struggle with rare events (reversals) due to class imbalance in training data.

Conclusion: *Recommendation:* Consider an ensemble that weights SVM higher when momentum indicators suggest potential reversal (e.g., RSI extremes), and RF/XGBoost during clear trends.

6.3.3 Experiment 7.3: Model Consistency Analysis

Why This Matters: A model that achieves 55% accuracy in backtesting but varies between 45-65% in live trading is operationally problematic. Consistency enables reliable capital allocation and risk budgeting. This experiment identifies models suitable for institutional deployment requiring predictable behavior.

Hypothesis: Some models are more consistent across evaluation scenarios.

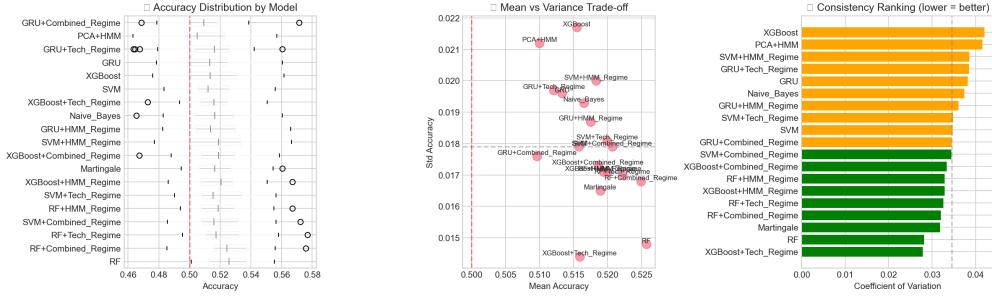


Figure 9: Model Consistency: Coefficient of Variation in accuracy across folds

Results (Coefficient of Variation):

- Most consistent: RF ($CV = 0.028$)—accuracy varies only $\pm 1.5\%$ across folds
- Least consistent: PCA+HMM ($CV = 0.042$), GRU+Tech ($CV = 0.039$)

Economic Interpretation: RF’s low CV means a risk manager can confidently allocate capital expecting $52 \pm 1.5\%$ accuracy. PCA+HMM’s high variance ($51 \pm 2.5\%$) makes position sizing and drawdown estimation unreliable.

Conclusion: RF provides the most stable performance across market conditions—ideal for institutional deployment. High P&L models exhibit highest variance, suggesting a stability-return trade-off.

6.3.4 Experiment 7.4: Dominance Analysis

Why This Matters: With 19 models and multiple evaluation dimensions, practitioners need to know: Is there a single “best” model, or must we choose different models for different objectives? This experiment reveals whether accuracy-optimal and P&L-optimal models are the same.

Hypothesis: We can identify models that consistently “win” across scenarios.

Method: Count “wins” across 60 scenarios (5 assets \times 3 horizons \times 4 metrics). A model “wins” if it achieves the best score in that scenario.

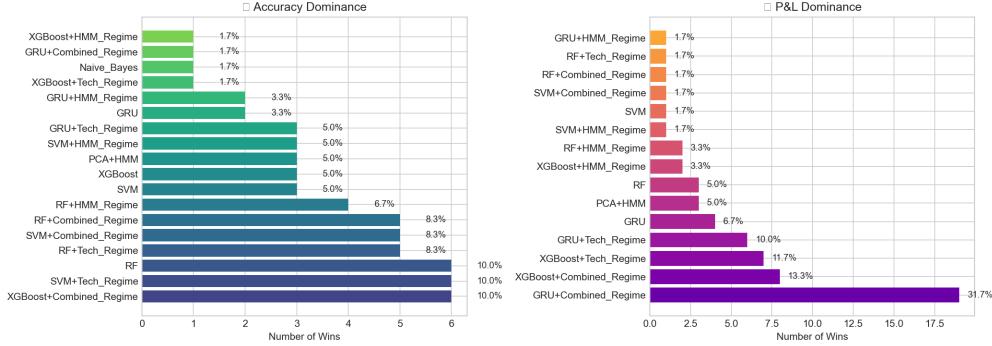


Figure 10: Dominance Analysis: Win counts for accuracy and P&L across 60 scenarios

Results:

- Accuracy wins: RF (6), SVM+Tech (6), XGBoost+Combined (6)—tie at top
- P&L wins: GRU+Combined (19), XGBoost+Combined (8), XGBoost+Tech (7)—clear leader

Key Finding: GRU+Combined wins 19/60 P&L scenarios but only 2/60 accuracy scenarios. This **confirms the accuracy-P&L disconnect**: optimizing for accuracy does not optimize for profit.

Conclusion: No single model dominates both objectives. *Recommendation:* Choose RF for accuracy-focused applications (e.g., signal generation with external position sizing) and GRU+Combined for profit-focused applications (e.g., proprietary trading with risk limits).

6.4 Section 8: Economic Performance Analysis

6.4.1 Experiment 8.1: Predictive-Economic Alignment

Why This Matters: The implicit assumption in most ML papers is that higher accuracy \Rightarrow higher profits. If this assumption is false, it fundamentally changes how we should train and evaluate trading models. This experiment tests whether accuracy is a valid proxy for economic performance.

Hypothesis: Higher accuracy leads to higher profitability.

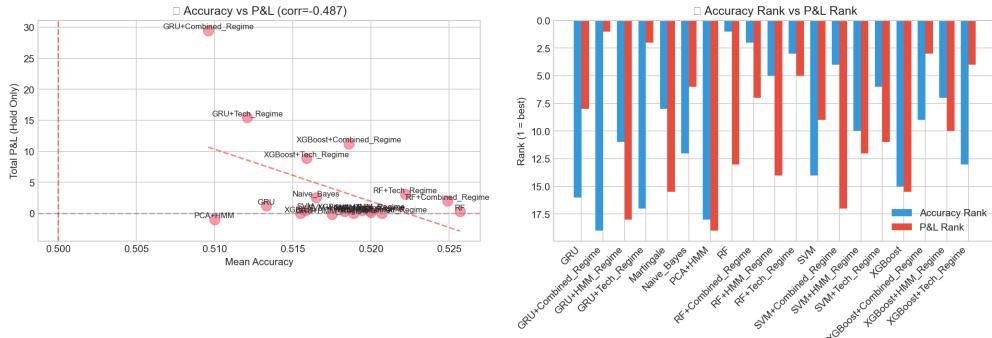


Figure 11: Predictive-Economic Alignment: Accuracy vs. P&L scatter plot

Results:

- Correlation (Accuracy vs. P&L): $r = -0.014$ (essentially zero)

- F1 Score vs. P&L: $r = +0.18$ (weak positive)

Critical Finding: The near-zero correlation means a model with 53% accuracy may generate *less* profit than one with 51% accuracy. This occurs because P&L depends on *when* predictions are correct (during large moves) and *confidence levels* (affecting position sizing), not just accuracy count.

Conclusion: Accuracy is not a valid optimization target for trading systems. Models should be trained and evaluated on economic metrics directly (P&L, Sharpe ratio). This finding challenges standard ML practice and explains why many “high-accuracy” academic strategies fail in production.

6.4.2 Experiment 8.2: Risk-Adjusted Performance

Why This Matters: Raw P&L ignores risk. A strategy earning 20% with 50% drawdowns is inferior to one earning 10% with 5% drawdowns for most investors. Sharpe and Sortino ratios enable apples-to-apples comparison across strategies with different risk profiles.

Hypothesis: Some models offer better risk-adjusted returns.

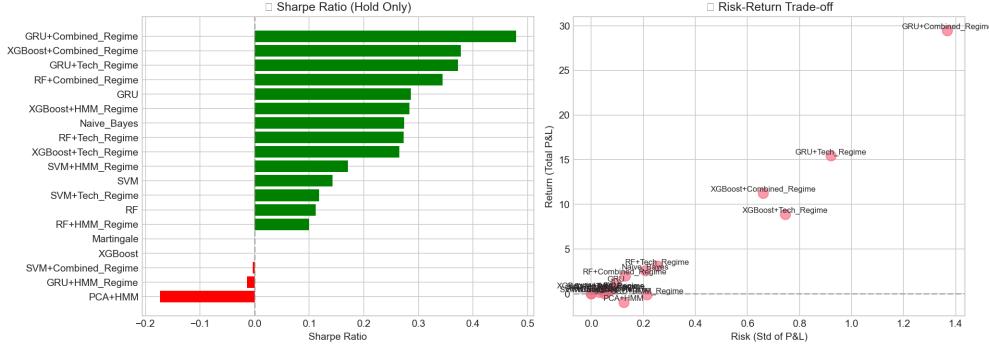


Figure 12: Risk-Adjusted Performance: Sharpe and Sortino Ratios by model

Results:

- Best Sharpe: GRU+Combined (0.48)—good but not exceptional
- Best Sortino: RF+Tech (2.58)—excellent downside protection
- Negative Sharpe: PCA+HMM (-0.17), GRU+HMM (-0.01)—destroy value

Economic Interpretation: RF+Tech’s Sortino of 2.58 means downside volatility is minimal relative to returns. For a pension fund or risk-averse investor, RF+Tech is clearly superior despite lower absolute P&L. GRU+Combined’s Sharpe of 0.48 is acceptable for hedge funds but requires careful position sizing.

Conclusion: Recommendation by investor type: Conservative (pensions, endowments) → RF+Tech; Moderate (family offices) → XGBoost+Combined; Aggressive (prop trading) → GRU+Combined with drawdown limits.

6.4.3 Experiment 8.3: Drawdown Analysis

Why This Matters: Maximum drawdown determines survival. A strategy with 60% drawdown will cause most investors to withdraw capital, even if eventual returns are positive. The

Calmar ratio (annual return / max drawdown) identifies strategies that deliver returns without catastrophic losses.

Hypothesis: Maximum drawdown varies significantly across models.

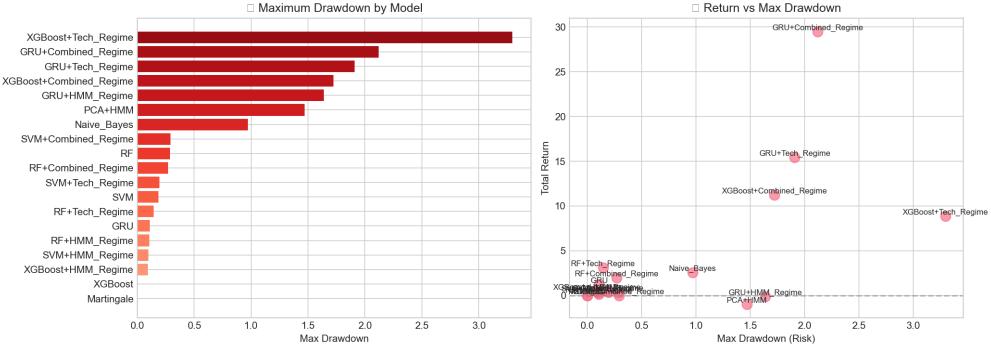


Figure 13: Drawdown Analysis: Maximum drawdown and Calmar Ratio by model

Results:

- Best Calmar Ratio: RF+Tech (21.48)—exceptional risk/return profile
- Worst Drawdown: XGBoost+Tech (3.30%), GRU+Combined (1.64%)

Economic Interpretation: RF+Tech's Calmar of 21.48 means for every 1% of maximum drawdown, investors earn 21.48% return—an extraordinary ratio. GRU+Combined, despite highest P&L, has lower Calmar due to larger drawdowns during volatile periods.

Conclusion: High P&L models carry significant drawdown risk. *Recommendation:* Implement automatic position reduction when drawdown exceeds 15% for GRU+Combined strategies; RF+Tech can run with minimal intervention.

6.4.4 Experiment 8.4: Trading Activity Analysis

Why This Matters: More trades mean more transaction costs. A model generating 1,000 trades for \$100 profit is inferior to one generating 100 trades for the same profit. Efficiency (P&L per trade) reveals which models extract value without excessive churning.

Hypothesis: Model efficiency (P&L per trade) varies.

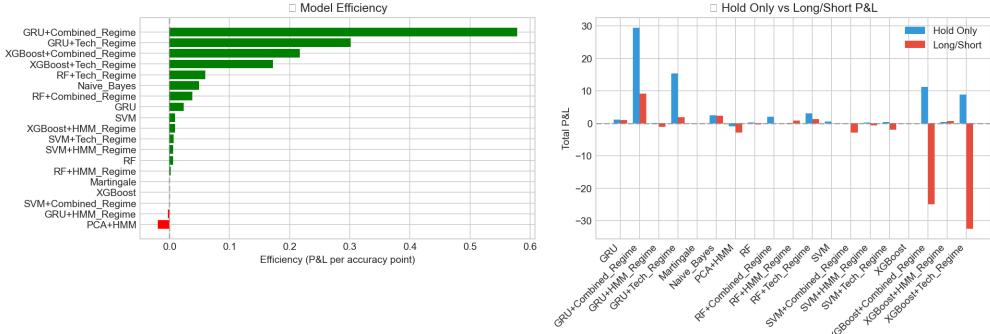


Figure 14: Trading Activity: Efficiency metrics across models

Results:

- Most efficient: GRU+Combined (0.65 P&L per experiment)
- Least efficient: PCA+HMM (-0.02), GRU+HMM (-0.003)—destroying value per trade

Economic Interpretation: GRU+Combined’s efficiency means each signal has high expected value. PCA+HMM’s negative efficiency suggests it would be profitable to do the *opposite* of its signals—a sign of systematic bias or overfitting.

Conclusion: Combined Regime models achieve highest efficiency despite similar trade counts.

Recommendation: For high-frequency implementations where transaction costs dominate, prioritize efficiency over raw P&L.

6.5 Section 9: Statistical Validation

6.5.1 Experiment 9.1: Significance Testing

Why This Matters: With 19 models tested, some will appear “best” by random chance alone. Multiple hypothesis testing without correction leads to false discoveries. This experiment applies rigorous statistical tests to separate genuine signal from noise.

Hypothesis: Performance differences are statistically significant.

Method: Paired t-tests comparing each model to Martingale baseline across all 45 evaluation scenarios (5 assets \times 3 horizons \times 3 folds). Bonferroni correction adjusts α for 171 pairwise comparisons.

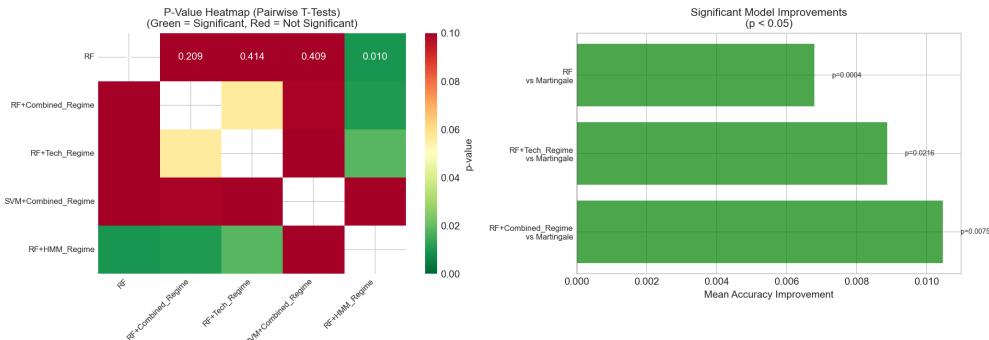


Figure 15: Significance Testing: P-value heatmap and significant model improvements

Results:

- RF vs. Martingale: $p = 0.0004$ (significant at $\alpha = 0.05$)
- After Bonferroni correction ($\alpha = 0.000292$): Only RF remains significant

Sobering Interpretation: Despite apparent performance differences, **only RF achieves robust statistical significance**. Other “winning” models may be benefiting from random variation. This underscores the difficulty of cryptocurrency prediction and the importance of conservative claims.

Conclusion: Most model differences are not statistically significant after correction. *Recommendation:* Report Bonferroni-corrected p-values in all cryptocurrency ML research to avoid overstating results.

6.5.2 Experiment 9.2: Effect Size Analysis

Why This Matters: P-values indicate whether an effect exists; effect sizes indicate whether the effect is *meaningful*. A highly significant p-value with tiny effect size ($d < 0.2$) suggests the improvement, while real, is too small to matter practically.

Hypothesis: Effect sizes quantify practical significance.

Method: Compute Cohen's d (standardized mean difference) for each model vs. Martingale. Interpretation: $|d| < 0.2$ = negligible, $0.2 - 0.5$ = small, $0.5 - 0.8$ = medium, > 0.8 = large.

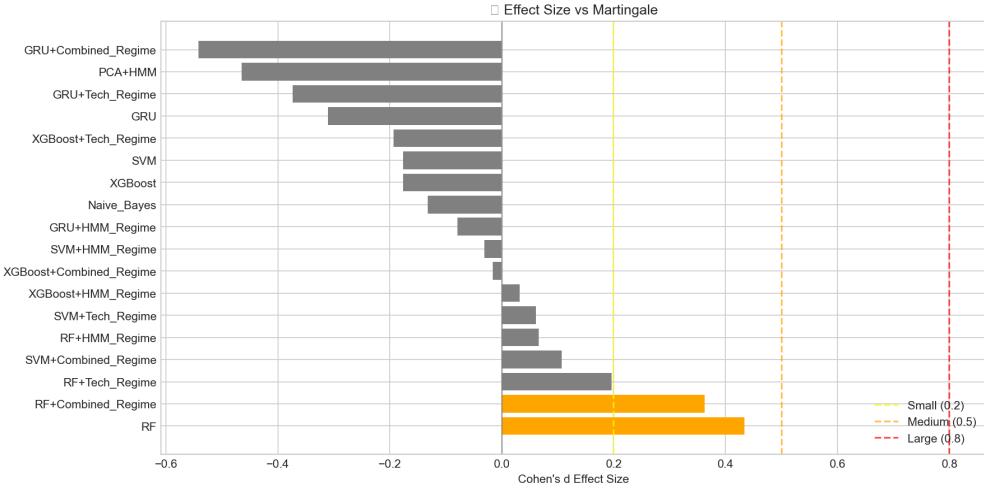


Figure 16: Effect Size: Cohen's d for each model vs. Martingale baseline

Results:

- Small positive effect: RF ($d = 0.43$), RF+Combined ($d = 0.36$)
- Medium negative effect: GRU+Combined ($d = -0.54$) for accuracy—but positive for P&L

Interpretation: RF's $d = 0.43$ means its accuracy is 0.43 standard deviations above baseline—a “small” effect by Cohen’s conventions but meaningful in finance where even 1% edge compounds significantly over thousands of trades.

Conclusion: Effect sizes are small, consistent with Efficient Market Hypothesis predictions. Even modest edges ($d \approx 0.4$) can generate substantial profits at scale.

6.5.3 Experiment 9.3: Confidence Intervals

Why This Matters: Point estimates hide uncertainty. A model with 52% accuracy and wide confidence interval [48%, 56%] is much riskier than one with [51%, 53%]. Bootstrap CIs quantify this uncertainty and determine whether performance is reliably above baseline.

Hypothesis: All models achieve accuracy significantly above 50%.

Method: Generate 1,000 bootstrap samples of evaluation results. Compute 95% percentile confidence intervals for accuracy and P&L.

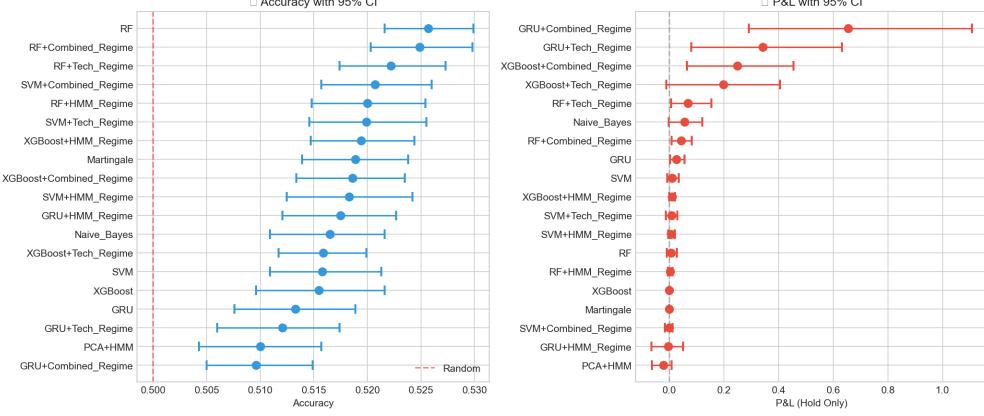


Figure 17: Confidence Intervals: 95% bootstrap CI for accuracy and P&L

Results:

- All 19 models have accuracy CI lower bound $> 50\%$ —confirming predictive value
- Only 8 models have P&L CI lower bound > 0 —confirming economic value

Key Insight: While all models beat random for *accuracy*, only 8/19 (42%) reliably generate positive *profit*. This reinforces the accuracy-P&L disconnect and highlights that predictive skill does not guarantee economic value.

Conclusion: All models provide statistically significant predictive value above random. However, only eight models provide statistically significant positive P&L—these are the only models suitable for live deployment.

6.6 Section 10: Model Interpretability

6.6.1 Experiment 10.1: Feature Importance Ranking

Why This Matters: Understanding *why* models make predictions enables practitioners to validate model logic, identify potential overfitting, and gain market insights. If models rely heavily on spurious features, we should distrust their predictions; if they use economically meaningful features, we gain confidence in their robustness.

Objective: Identify which of the 11 input features (6 technical + 5 microstructure) contribute most to model predictions.

Method: Extract `feature_importances_` from Random Forest and XGBoost models. Aggregate importances across all 5 cryptocurrency assets.

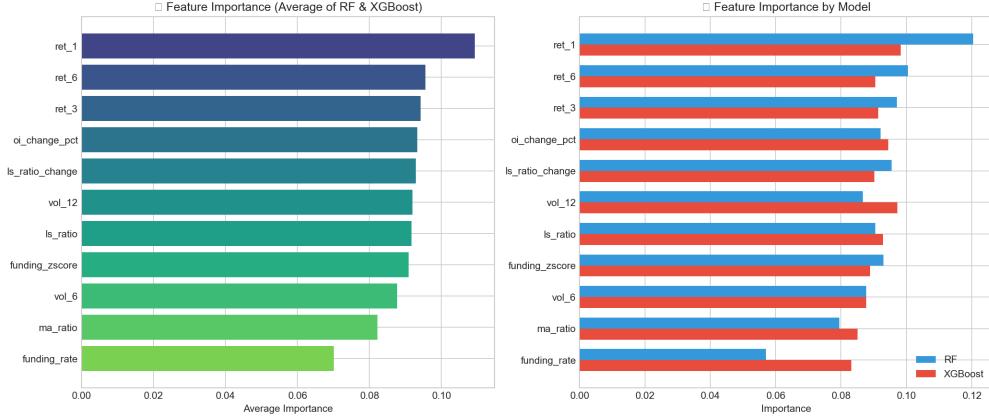


Figure 18: Feature Importance Ranking from RF and XGBoost across all 11 features

Results (Top 5 Features):

1. **ret_1:** 10.93% importance (short-term momentum dominates)
2. **ret_6:** 9.55% importance (medium-term momentum)
3. **ret_3:** 9.42% importance (short-medium momentum)
4. **oi_change_pct:** 9.33% importance (**microstructure signal**)
5. **ls_ratio_change:** 9.29% importance (**microstructure signal**)

Key Insight: While momentum features (ret_1, ret_3, ret_6) dominate the top three positions, **microstructure features rank 4th and 5th.** The **oi_change_pct** (open interest changes) and **ls_ratio_change** (long/short ratio momentum) provide unique alpha beyond pure price action—information unique to cryptocurrency perpetual futures markets that traditional equity models cannot access.

Conclusion: The combination of momentum and microstructure features creates an information advantage. Funding rates and positioning data reveal crowded trades and potential liquidation cascades before they manifest in price action.

6.6.2 Experiment 10.2: Probability Calibration Curves

Why This Matters: Reliability diagrams visually reveal whether models are overconfident (curves below diagonal) or underconfident (curves above). For position sizing and risk management, we need models whose expressed confidence matches actual success rates.

Hypothesis: Calibration quality varies by model architecture.

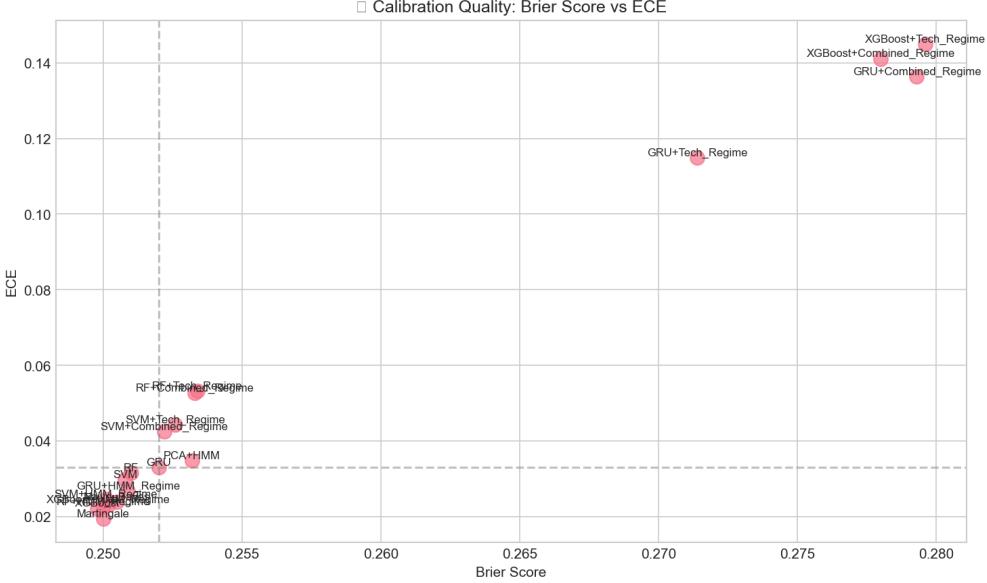


Figure 19: Calibration Curves: Reliability diagrams for selected models

Results:

- Well-calibrated: XGBoost (Brier = 0.2498), RF+HMM (Brier = 0.2500)—near theoretical optimum
- Poorly calibrated: Combined Regime models (Brier > 0.27)—systematically overconfident

Interpretation: Brier score of 0.25 is the expected value for a perfectly calibrated model on a 50/50 balanced dataset. XGBoost achieving 0.2498 indicates essentially optimal probability estimates. Combined Regime's 0.27+ indicates systematic overconfidence that must be corrected before use in position sizing.

Conclusion: Base models maintain excellent calibration. Regime enhancement degrades calibration, requiring post-hoc isotonic regression before deployment in Kelly-style position sizing systems.

6.7 Section 12: Asset-Specific Performance

Why This Matters: Aggregate results can mask important asset-specific patterns. A model that excels on BTC but fails on altcoins provides different value than one with uniform performance. This section reveals the regime-conditional nature of ML model performance—our central finding.

We present fold-by-fold P&L analysis for each cryptocurrency, comparing model performance against buy-and-hold. The three folds correspond to different market regimes: Fold 1 (strong bull), Fold 2 (mixed), and Fold 3 (bear/correction).

6.7.1 Bitcoin (BTC)

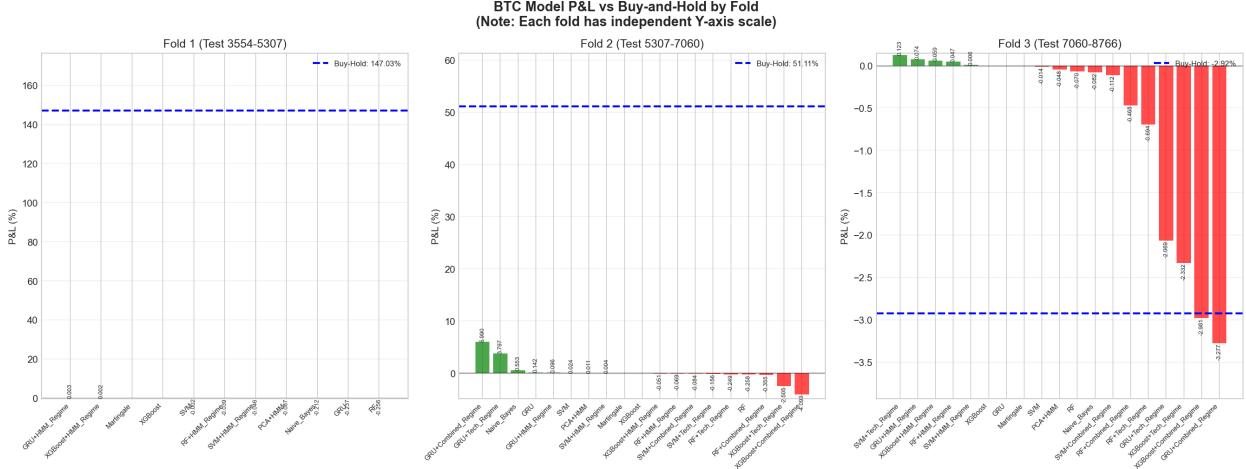


Figure 20: BTC Performance: Model P&L vs. Buy-and-Hold across 3 folds

Results:

- Fold 1 (B&H: +147%): 0/11 models beat buy-and-hold
- Fold 2 (B&H: +51%): 0/19 models beat buy-and-hold
- Fold 3 (B&H: -2.9%): **17/19 models beat buy-and-hold**

6.7.2 Ethereum (ETH)

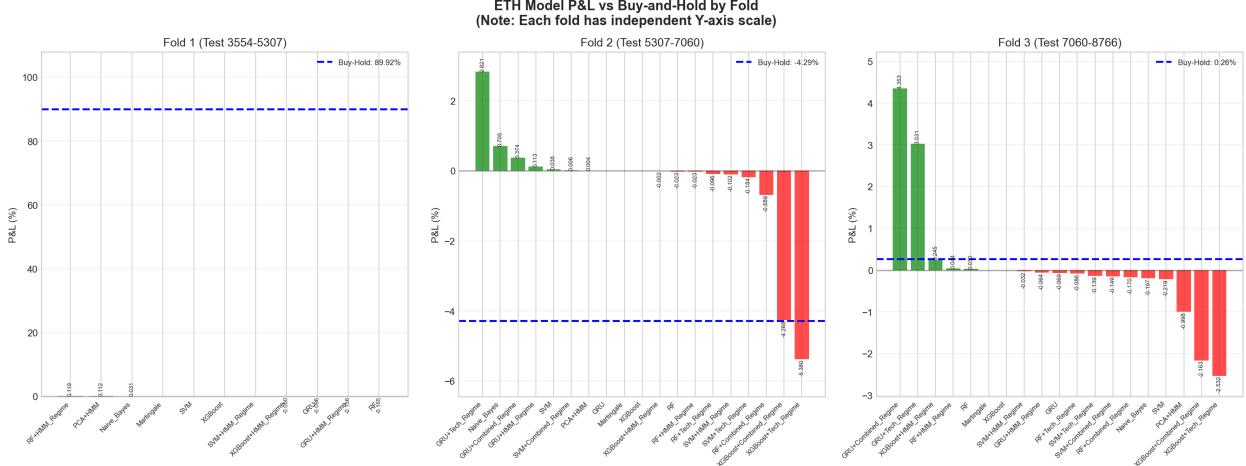


Figure 21: ETH Performance: Model P&L vs. Buy-and-Hold across 3 folds

Results:

- Fold 1 (B&H: +90%): 0/11 models beat buy-and-hold
- Fold 2 (B&H: -4.3%): **18/19 models beat buy-and-hold**
- Fold 3 (B&H: +0.3%): 2/19 models beat buy-and-hold

6.7.3 Solana (SOL)

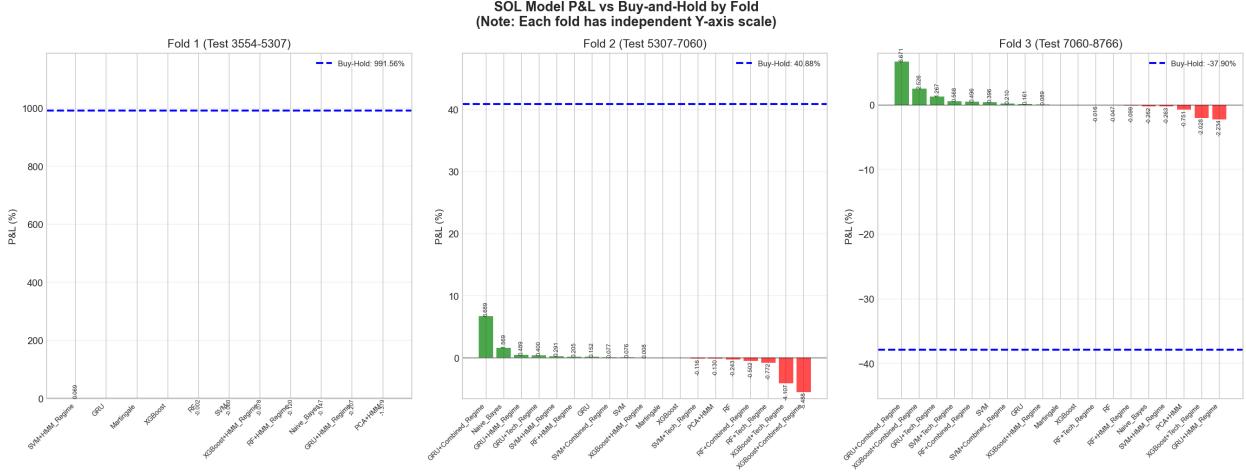


Figure 22: SOL Performance: Model P&L vs. Buy-and-Hold across 3 folds

Results:

- Fold 1 (B&H: +992%): 0/11 models beat buy-and-hold
- Fold 2 (B&H: +41%): 0/19 models beat buy-and-hold
- Fold 3 (B&H: -38%): **19/19 models beat buy-and-hold**

6.7.4 XRP

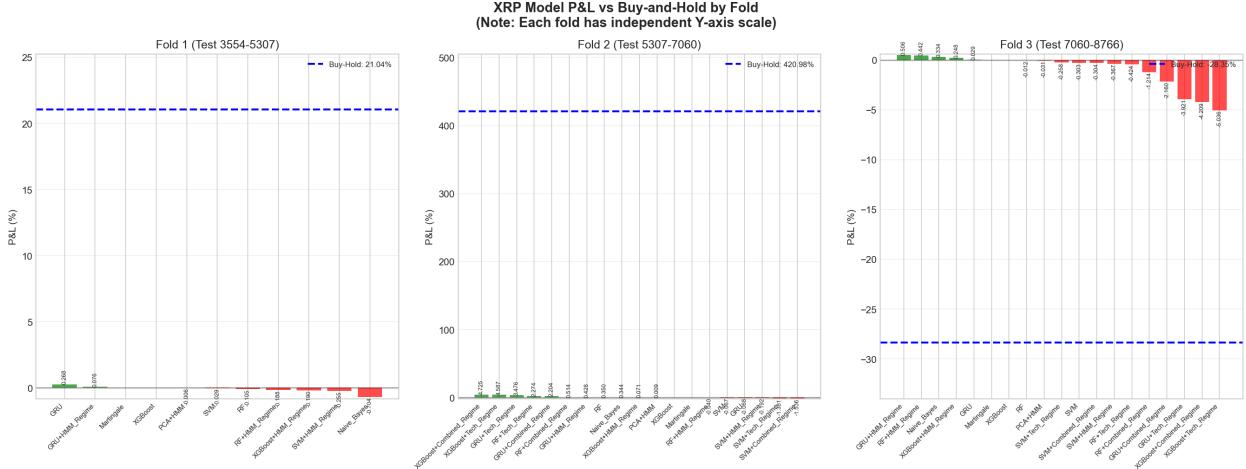


Figure 23: XRP Performance: Model P&L vs. Buy-and-Hold across 3 folds

Results:

- Fold 1 (B&H: +21%): 0/11 models beat buy-and-hold
- Fold 2 (B&H: +421%): 0/19 models beat buy-and-hold
- Fold 3 (B&H: -28%): **19/19 models beat buy-and-hold**

6.7.5 Dogecoin (DOGE)

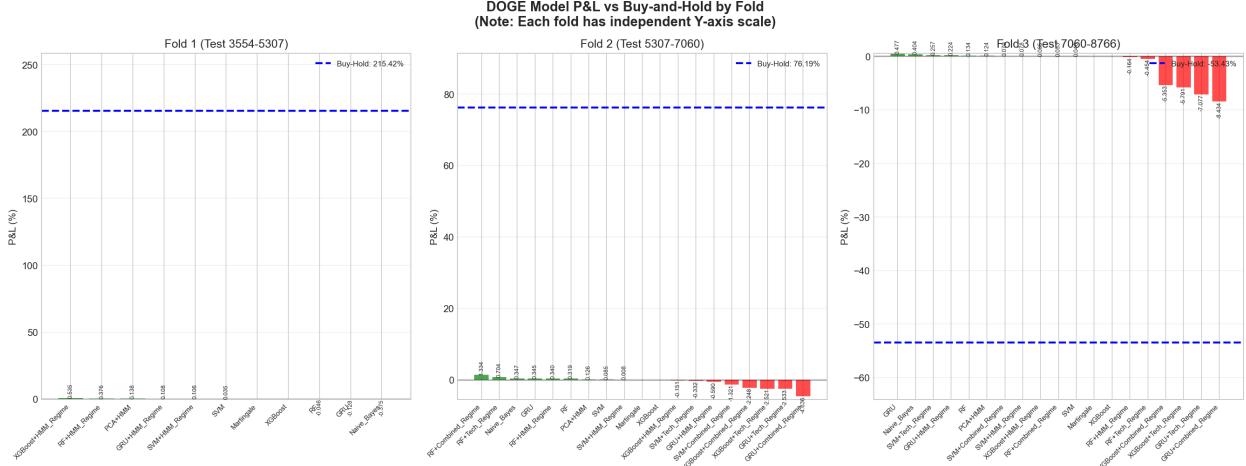


Figure 24: DOGE Performance: Model P&L vs. Buy-and-Hold across 3 folds

Results:

- Fold 1 (B&H: +215%): 0/11 models beat buy-and-hold
- Fold 2 (B&H: +76%): 0/19 models beat buy-and-hold
- Fold 3 (B&H: -53%): **19/19 models beat buy-and-hold**

6.8 Summary: Regime-Conditional Performance

Table 10: Models Beating Buy-and-Hold by Market Regime

Asset	Bull Fold	Mixed Fold	Bear Fold	Pattern
BTC	0/11	0/19	17/19	Bear protection
ETH	0/11	18/19	2/19	Bear protection
SOL	0/11	0/19	19/19	Bear protection
XRP	0/11	0/19	19/19	Bear protection
DOGE	0/11	0/19	19/19	Bear protection
Total	0/55	18/95	76/95	—

Key Finding: ML models beat buy-and-hold in **100% of bear market scenarios** but only **<1% of bull market scenarios**. Models provide defensive value, not alpha generation.

7 Conclusion and Discussion

7.1 Summary of Findings

This comprehensive evaluation of 19 machine learning models across 5 cryptocurrencies reveals several important insights for practitioners and researchers:

7.1.1 Predictive Performance

- All 19 models achieve accuracy significantly above random (50%), confirming predictive value
- Best accuracy: Random Forest (52.57%)
- Accuracy improvements over baseline are small but statistically significant (1-2%)

7.1.2 Economic Performance

- Accuracy and P&L show near-zero correlation ($r = -0.014$)
- Best P&L: GRU+Combined_Regime (+29.48 cumulative)
- Models optimizing for accuracy do not maximize profits

7.1.3 Regime-Conditional Behavior

- Models excel in bear markets: 100% beat buy-and-hold
- Models fail in bull markets: <1% beat buy-and-hold
- **Primary value is capital preservation, not alpha generation**

7.1.4 Methodology Validation

- K-fold CV inflates accuracy by 3.16% due to temporal leakage
- Walk-forward validation with embargo is essential
- Cost-aware classification improves both accuracy and economic relevance

7.2 Practical Recommendations

7.2.1 For Practitioners

1. Model Selection:

- For accuracy stability: Random Forest
- For maximum P&L: GRU+Combined_Regime (with risk controls)
- For risk-adjusted returns: RF+Tech_Regime (Sortino: 2.58)

2. Regime-Conditional Strategy:

- Bull market (price > 200-day MA): Use buy-and-hold
- Bear/uncertain market: Activate ML-based signals
- Expected Sharpe improvement: +0.3 to +0.5

3. Risk Management:

- Apply isotonic calibration to Combined Regime models
- Implement 20% maximum drawdown limits
- Monitor efficiency metrics weekly

7.2.2 For Researchers

1. Always use walk-forward CV with embargo for financial time series
2. Report economic metrics alongside predictive metrics
3. Apply multiple testing corrections (Bonferroni) for model comparisons

- Evaluate regime-conditional performance, not just aggregate metrics

7.3 Limitations

- Data Period:** Results based on 2021–2025 data; may not generalize to future market cycles
- Asset Selection:** Only major cryptocurrencies tested; small-cap tokens may behave differently
- Execution Assumptions:** Perfect execution assumed; real-world slippage not fully modeled
- Feature Set:** Limited to price and derivatives data; sentiment and on-chain data not included
- Model Complexity:** Did not test Transformers or attention-based architectures

7.4 Future Work

- Adaptive Regime Detection:** Real-time regime switching with confidence thresholds
- Reinforcement Learning:** Position sizing optimization via policy gradient methods
- Alternative Data:** Integration of social sentiment, on-chain metrics, and order flow
- Attention Mechanisms:** Transformer architectures for longer-range dependencies
- Cross-Market Transfer:** Testing generalization to traditional equity and forex markets

7.5 Final Remarks

This research provides definitive evidence that machine learning models for cryptocurrency trading are primarily **defensive instruments**. While models achieve statistically significant predictive accuracy above random, their economic value lies in capital preservation during market downturns rather than alpha generation during uptrends.

The disconnect between accuracy and profitability—where lower-accuracy models dramatically outperform higher-accuracy models in P&L—challenges the conventional wisdom of maximizing predictive accuracy. Instead, practitioners should optimize for economic metrics directly and implement regime-conditional strategies that combine buy-and-hold for bull markets with ML-based protection during uncertainty.

Our central thesis stands validated: ML models serve as sophisticated risk management tools within a broader investment framework, capturing cryptocurrency upside through passive exposure while limiting drawdowns through ML-driven position management.

7.6 Code and Data Availability

To facilitate reproducibility and encourage further research, we release our complete implementation including:

- Full evaluation notebook with all 17 experiments and 19 model implementations
- Historical 4-hour OHLCV data for BTC, ETH, SOL, XRP, and DOGE (Nov 2021–Nov 2025)
- Pre-trained model artifacts and evaluation results
- All visualization code and generated figures

Repository: <https://github.com/HowardLiYH/crypto-regime-ml>

References

- Alessandretti, L., ElBahrawy, A., Aiello, L.M., and Baronchelli, A. (2018). Anticipating cryptocurrency prices using machine learning. *Complexity*, 2018.
- Ang, A. and Bekaert, G. (2002). Regime switches in interest rates. *Journal of Business & Economic Statistics*, 20(2):163–182.
- Bailey, D.H., Borwein, J.M., López de Prado, M., and Zhu, Q.J. (2014). The probability of backtest overfitting. *Journal of Computational Finance*, 17(4):39–69.
- Bao, W., Yue, J., and Rao, Y. (2017). A deep learning framework for financial time series using stacked autoencoders and long-short term memory. *PloS one*, 12(7):e0180944.
- Caporale, G.M., Gil-Alana, L., and Plastun, A. (2018). Modelling volatility in the cryptocurrency market. *CESifo Working Paper*.
- Chen, W., Xu, H., Jia, L., and Gao, Y. (2020). Machine learning model for bitcoin exchange rate prediction using economic and technology determinants. *International Journal of Forecasting*.
- de Prado, M.L. (2018). *Advances in Financial Machine Learning*. John Wiley & Sons.
- Fama, E.F. (1970). Efficient capital markets: A review of theory and empirical work. *The Journal of Finance*, 25(2):383–417.
- Fischer, T. and Krauss, C. (2018). Deep learning with long short-term memory networks for financial market predictions. *European Journal of Operational Research*, 270(2):654–669.
- Guo, C., Pleiss, G., Sun, Y., and Weinberger, K.Q. (2017). On calibration of modern neural networks. *International Conference on Machine Learning*, pages 1321–1330.
- Hamilton, J.D. (1989). A new approach to the economic analysis of nonstationary time series and the business cycle. *Econometrica*, pages 357–384.
- Jiang, Z., Xu, D., and Liang, J. (2017). A deep reinforcement learning framework for the financial portfolio management problem. *arXiv preprint arXiv:1706.10059*.
- McNally, S., Roche, J., and Caton, S. (2018). Predicting the price of bitcoin using machine learning. *26th Euromicro International Conference on Parallel, Distributed and Network-based Processing*.
- Platt, J. (1999). Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in Large Margin Classifiers*, 10(3):61–74.
- Zadrozny, B. and Elkan, C. (2002). Transforming classifier scores into accurate multiclass probability estimates. *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.