Comparative Evaluation of ML Models for Crypto Signal Generation:

A Walk-Forward Analysis with Regime Enhancement

CIS 5200 Presentation

Howard Li: Pre-processing, workflow design, and evaluations

Nitin Lodha: Model generation and evaluations

Akshat Bokdia: Model generation and evaluations

Motivation: Addressing the Challenges in Crypto Signal Generation

1.1 The Crypto Challenge

- Extreme volatility, 24/7 trading, rapid information flow.
- Regime switches transform market dynamics in hours.
- Short-horizon forecasting is technically demanding but valuable.

1.2 The Fallacy of Point Estimates

- Single number predictions have limited actionable value.
- Probability is crucial for effective position sizing (e.g., Kelly Criterion).
- Uses probability calibration for accurate confidence.

1.3 Paper vs. Realized Returns

- Backtests often ignore real-world frictions.
- Frictions: Slippage, spread widening, transaction costs.
- Implements cost-aware classification (P(return > cost threshold)).

1.4 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?

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
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)

Feature Definitions

Base Technical Features (6 features)

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)

Table: Base Technical Feature Definitions

Feature	Description	Calculation
ret 1	1-bar log return	$\log(P_t/P_{t-1})$
ret 3	3-bar log return	$\log(P_t/P_{t-3})$
ret 6	6-bar log return	$\log(P_t/P_{t-6})$
vol 6	6-bar volatility	$\mathrm{std}(\mathrm{ret}\ 1)$ over 6 bars
vol 12	12-bar volatility	$\mathrm{std}(\mathrm{ret}\ 1)$ over 12 bars
ma ratio	MA crossover ratio	$\log(\mathrm{MA}_{10}/\mathrm{MA}_{20})$

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: Microstructure Feature Definitions

Feature	Description	Calculation
funding rate	Perpetual funding rate	Raw 8-hour funding rate
funding zscore	Standardized funding	$FR_t - \mu_F R_{,50} / \sigma_{FR,50}$
ls ratio	Long/Short ratio	Long Positions / Short Positions
ls ratio change	L/S ratio momentum	3-bar percentage change in L/S ratio
oi change pct	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.

Methods: Regime Features

HMM Regime Features

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

$$\text{regime features} = \begin{bmatrix} P(S_t = 0), \\ P(S_t = 1), \\ P(S_t = 2)] \end{bmatrix} \tag{15}$$

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

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

Combined Regime Features

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

Base Models

1 Random Forest (RF)

Random Forest is an ensemble of decision trees trained on bootstrap samples with random feature subsets. We use 100 trees with maximum depth of 10 and balanced class weights:

$$\boxed{\hat{y} = \mathrm{mode}\left(\{h_b(\mathbf{x})\}_{b=1}^{B}\right)} \tag{6}$$

where h_b is the b-th tree and $B = 100$.

2 Support Vector Machine (SVM)

We employ SVM with RBF kernel and probability calibration via Platt scaling:

$$\boxed{f(\mathbf{x}) = \mathrm{sign}\left(\sum_{i \in SV} \alpha_i y_i K(\mathbf{x}_i, \mathbf{x}) + b\right)} \tag{7}$$

where $K(\mathbf{x}_i, \mathbf{x}) = \exp(-\gamma \|\mathbf{x}_i - \mathbf{x}\|_2^2)$ is the RBF kernel.

3 XGBoost

XGBoost implements gradient boosting with regularization:

$$\boxed{\hat{y}^{(t)}{}_i = \hat{y}^{(t-1)}{}_i + \eta f_t(\mathbf{x}_i)} \tag{8}$$

where η is the learning rate and f_t is the t-th tree. We use 100 estimators with learning rate 0.1 and maximum depth 5.

4 Gated Recurrent Unit (GRU)

GRU is a recurrent architecture that processes sequences of features:

$$\boxed{\begin{aligned}
z_t &= \sigma(W_z \mathbf{x}_t + U_z h_{t-1}) && (9)\\
r_t &= \sigma(W_r \mathbf{x}_t + U_r h_{t-1}) && (10)\\
\tilde{h}_t &= \tanh(W_h \mathbf{x}_t + U_h (r_t \odot h_{t-1})) && (11)\\
h_t &= (1 - z_t) \odot h_{t-1} + z_t \odot \tilde{h}_t && (12)
\end{aligned}}$$

We use a sequence length of 20 bars, 64 hidden units, and dropout of 0.2.

PCA + HMM and Benchmark Models

5 PCA + Hidden Markov Model (PCA+HMM)

This model combines dimensionality reduction with regime detection:

1. Apply PCA to reduce features to principal components
2. Fit Gaussian HMM to identify latent market states
3. Compute state probabilities for classification

The HMM defines:

$$\boxed{P(\mathbf{X}_t|S_t = j) = \mathcal{N}(\mathbf{X}_t; \mu_j, \Sigma_j)} \tag{13}$$

$$\boxed{P(S_t = j|S_{t-1} = i) = A_{ij}} \tag{14}$$

where S_t is the hidden state, $\mathbf{A}$ is the transition matrix, and (μ_j, Σ_j) are state-dependent emission parameters.

Benchmark Models

1 Naive Bayes

Gaussian Naive Bayes assumes feature independence:

$$\boxed{P(\mathbf{y}|\mathbf{x}_1, ..., \mathbf{x}_n) \propto P(\mathbf{y}) \prod_{i=1}^{n} P(\mathbf{x}_i|\mathbf{y})} \tag{16}$$

2 Martingale

The martingale benchmark predicts no change:

$$\boxed{\hat{p} = 0.5 \quad \forall t} \tag{17}$$

This represents the efficient market hypothesis baseline.

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

Walk-Forward Data Split

We implement expanding window walk-forward cross-validation with three folds:

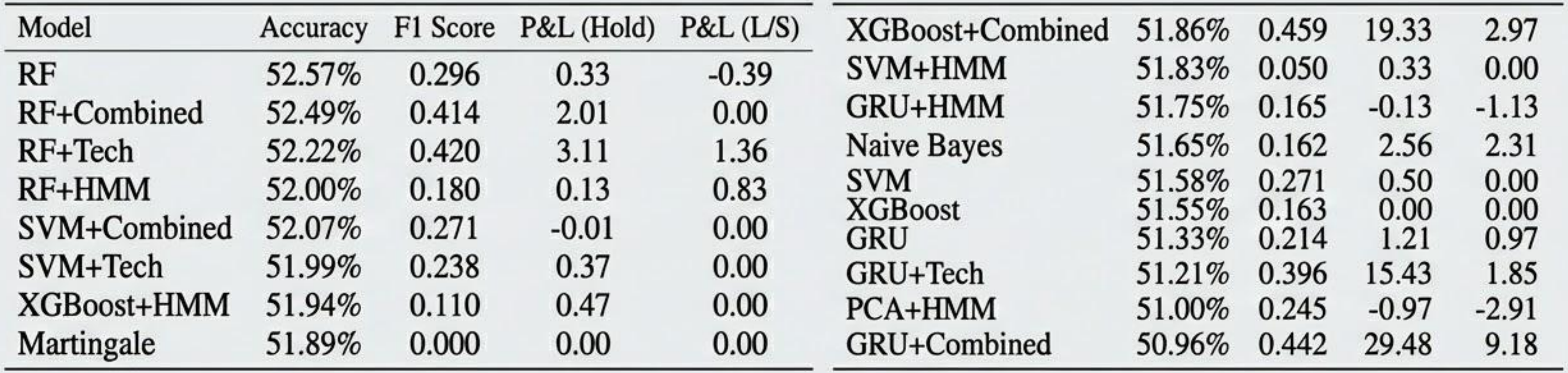
Table 4: Walk-Forward Split Structure

Fold	Training Period	Test Period	Test Duration
1	Nov 2021 – Aug 2022	Jun 2023 – Apr 2024	$\sim$10 months
2	Nov 2021 – Jun 2023	Apr 2024 – Jan 2025	$\sim$10 months
3	Nov 2021 – Apr 2024	Jan 2025 – Nov 2025	$\sim$10 months

A 24-bar (96-hour) embargo period separates training and test sets to prevent temporal autocorrelation leakage.

Main Evaluation Results

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



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).

Regime Feature Comparison

Hypothesis: Regime features improve model performance.

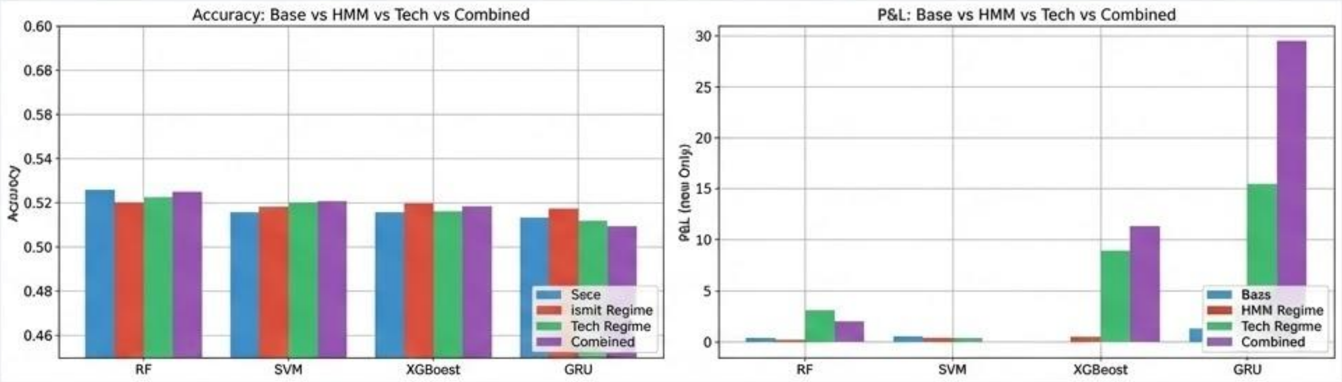


Figure: Regime Comparison: Base vs. HMM vs. Technical vs. Combined regime features

Results:

- HMM Regime: Modest accuracy improvement, stable calibration
- Technical Regime: Moderate P&L improvement
- Combined Regime: Highest P&L potential, poorest calibration

Conclusion: Regime features provide P&L improvement at the cost of calibration quality. Combined regime offers highest reward with highest variance.

Regime-Conditional Performance: ML Models vs. Buy-and-Hold

- 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

Table 8: 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.

Drawdown Analysis

Hypothesis: Maximum drawdown varies significantly across models.

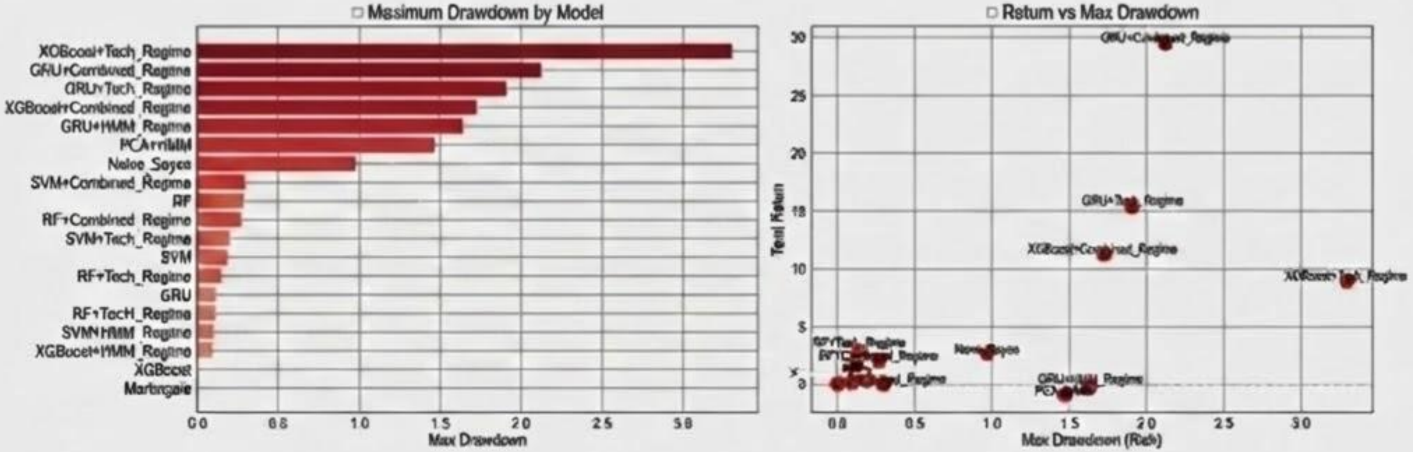


Figure 13: Drawdown Analysis: Maximum drawdown and Calmar Ratio by model

Results:

- Best Calmar Ratio: RF+Tech (21.48)
- Worst Drawdown: XGBoost+Tech (3.30), GRU+Combined (1.64)

Conclusion: High P&L models carry significant drawdown risk. RF+Tech balances return and drawdown most effectively.

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