Cryptocurrency Price Forecasting: Benchmark Results for Simple Statistical Models

Yu Li* Yusong Duan Haowen Wu

June 16, 2025

Contents

1 Abstract

(This is a handout version of the slides not final report)

This project presents an empirical evaluation of cryptocurrency price forecasting models using daily market data. We implement and compare ARIMA, MLP, RNN, LSTM and BiLSTM models with rigorous statistical test and financial back-testing. The results show that Bi-LSTM model outperforms other and we also conduct multiple robustness checks to ensure models' realistic application.

Note: Notebook is achieved in Kaggle, you may refer to https://www.kaggle.com/code/liyushi669/cryptocoin-baseline-model. For Replication,Pandas (2.2.3),Numpy(1.25.0),Torch(1.9.1+cpu) and Python(3.9.19) with random seeds quoted in the notebook are required.

2 Introduction

2.1 Context

Cryptocurrency markets have emerged as a novel asset class with distinct characteristics that challenge traditional financial modeling approaches.

- Market Microstructure
  - 24/7 trading with no market closures
  - Decentralized exchange ecosystems
  - High-frequency algorithmic trading dominance
- Statistical Properties
  - Excess volatility (10x equities market)
  - Non-normal return distributions
  - Time-varying volatility clusters

- Economic Significance
  - $1.2 trillion total market capitalization (2024)
  - Growing institutional adoption
  - Hedge against fiat currency risks This unique combination of features necessitates specialized forecasting methodologies beyond conventional time-series approaches.

2.2 Research Question

Our study establishes a comprehensive benchmark for cryptocurrency price forecasting by addressing:

- Baseline Performance
  - How do classical and modern approaches compare as foundational baselines?
  - What metrics form the essential evaluation framework?
- Implementation Standards
  - What are the reproducible implementation standards for each model class?
  - How to establish fair comparison protocols?
- Extension Pathways
  - What are the most promising directions for model enhancement?
  - Where do current approaches show limitations?

3 Methodology & Data

3.1 Data Sources

The analysis uses top 50 trading volume up to March 2024, daily cryptocurrency market data form CoinGecko API.

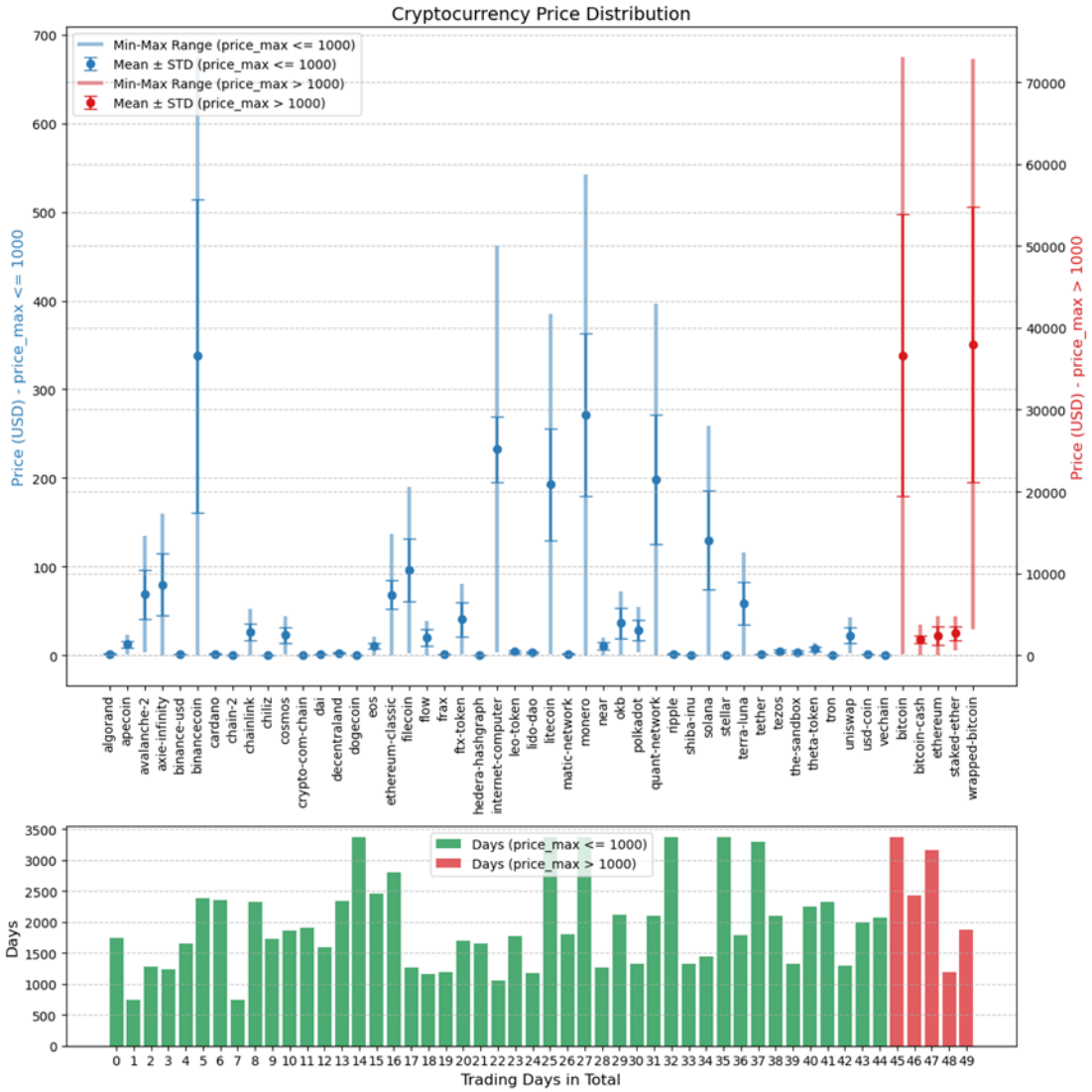

Figure 1: Cryptocurrency Price and Volume Distribution

3.2 Model Implementation

- Classical approaches:ARIMA
- Modern approaches(Simple DL):
  - MLP (2 layers)
  - RNN (2 layers)
  - LSTM (single layer)
  - BILSTM (single layer)

3.3 Evaluation Metrics

3 Cat. metrics:

- Stat. Metrics: NRMSE,WMAPE,MAPE,R-squared

- Fin. Metrics: Sharpe Ratio, Accuracy, Precision,F1-Score,MDD
- Comp. Metrics: Training & Inference (Not covered yet)

4 Background

4.1 Introduction to Crypto-coins

Crypto-coins are digital tokens that are created and issued by crypto-currencies like Bitcoin, Ethereum, and Litecoin. They are used to store and transact digital assets, and they are the primary form of payment for most digital services.

4.2 Why choose?

The main reason why we choose crypto-coins is that they are relatively new and have a low barrier to entry. You can buy and sell them easily, and they are not affected by traditional financial institutions like banks and credit unions.

4.3 Why matters?

1. Mis-priced Assets
2. High Volatility
3. High Unpredictability
4. No Traditional Financial Institutions

4.4 Descriptive Statistics

4.4.1 50 coin correlation matrix

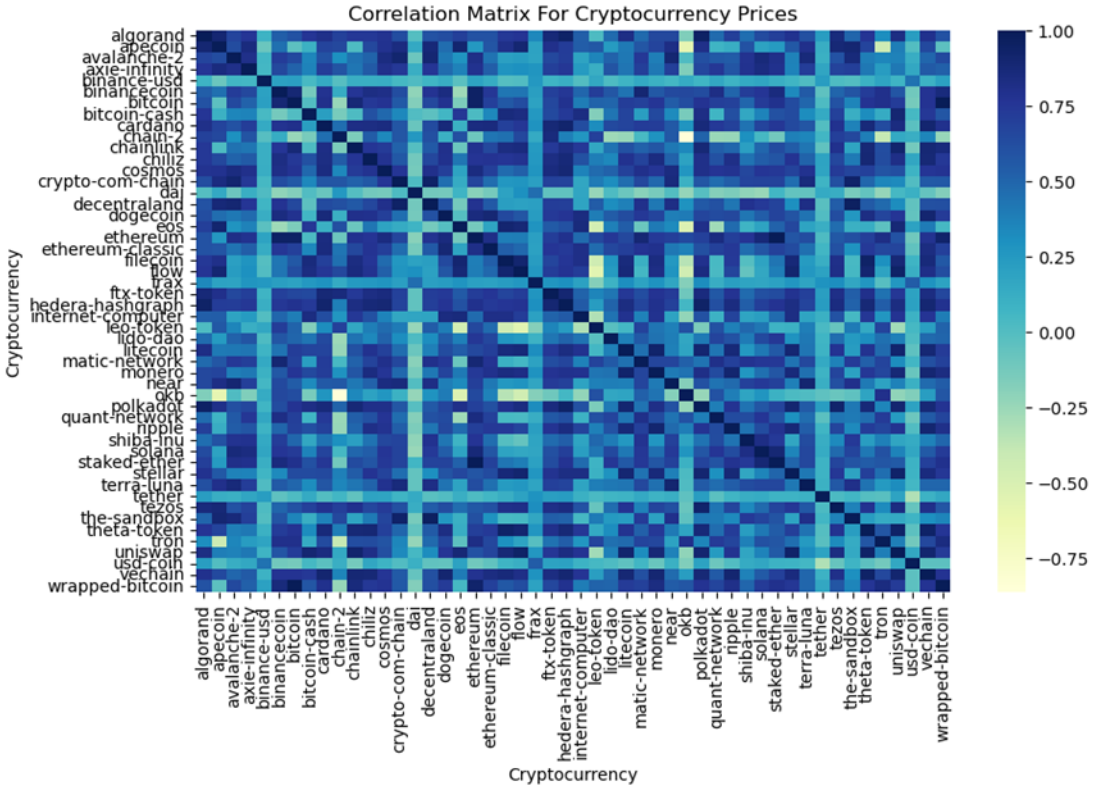

Figure 2: High Correlation Matrix

4.4.2 Single coin Data Sample

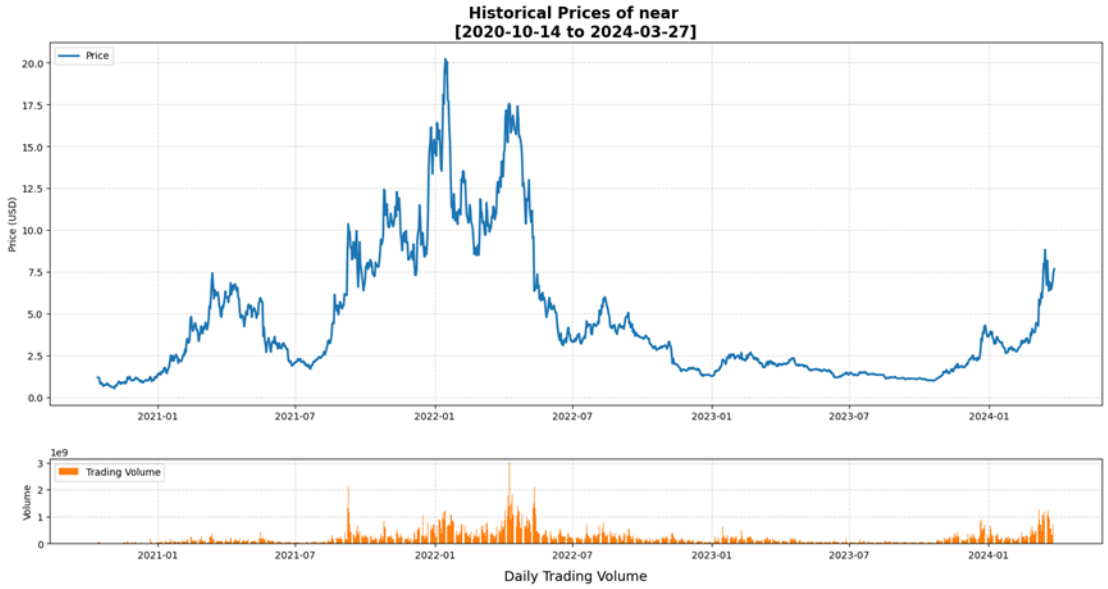

Figure 3: High Volatility Coin

5 Methodology

5.1 Preset Env.

- fixed train-validation-test split (64-16-20);
- Identical computational environment random seeds 2025 for cuda, all models execute on hardware/software identical platforms to eliminate performance bias from computational resource disparities.
- fixed other parameter like learning rate, layers and neurons, etc.

5.2 Hyperparameter Tuning

Training Loss Function: MSE

Hyperparameter Type	Candidate Values
Activation Function	relu, elu, sigmoid, tanh
Optimizer	adam, sgd, rmsprop
Epochs	10, 20, 40, 80, 160
Batch Size	8, 16, 32, 64, 128

5.3 Evaluation Metrics

5.3.1 Statistical Metrics

Metrics:

- NRMSE(Normalized RMSE)
- WMAPE(Weighted Mean Absolute Percentage Error)
- MAPE(Mean Absolute Percentage Error)
- R-squared

For reference only, we will know why this is the case.

5.3.2 Financial Metrics(Simulated Trading)

Metrics: Accuracy, Precision, Recall, F1-Score, MDD, Sharpe Ratio,Avg. Return, Total Return

F1 Score is precision-oriented metrics which means $\beta == 0.5$:

$$F1\_score = \frac{(1+\beta^2) \times (precision \times recall)}{\beta^2 \times precision + recall}$$

Signal Generation Process by LSTM with 7-day lagged coin "near"

Figure 4: Signal Sample Graph

6 Results

6.1 Model Performance Results

6.1.1 For Coin "NEAR" Financial Performance

NEAR	Win Ratio	Precision	F1 Score	Avg Return	Sharpe Ratio	Max Drawdown	Total Return
ARIMA	0.3636	0.2731	0.2874	0.0633	0.9611	-0.3332	0.2764
MLP	0.3594	0.1302	0.1492	0.3323	1.3018	-0.3002	0.3323
RNN	0.3208	0.2902	0.2958	0.0541	0.9975	-0.3002	0.4390
LSTM	0.3042	0.3469	0.3374	0.6081	1.9762	-0.3002	1.9542
BILSTM	0.3906	0.3971	0.3958	0.0954	2.3099	-0.2294	0.5677

BiLSTM have the best overall performance because of its 1) Highest prediction quality, i.e., Win Ratio (39.06%), Precision (39.71%), F1 Score (39.58%) all rank 1st among models; 2) Optimal risk control because it has the smallest Max Drawdown(-22.94%), which is 10% lower than worst-performing ARIMA; 3) Elite risk-adjusted returns, Sharpe Ratio (2.31) is the highest overall.

LSTM have the highest profitability but together with volatility. It has exceptional profit generation with avg return of 60.81% and total return of 195.42% which outperforms all other models by wide margins, however, it also has high-risk exposure, since its max drawdown is -30.02% which is the near-worst level and win ratio is 30.42%, showing lowest accuracy and correct in only 3/10 trades. Further, we find that the 195% total return is actually from capturing extreme rallies (e.g., NEAR's +300% surge in 2023), which means its profitability relies on capturing extreme price movements.

6.1.2 For Coin "NEAR" Statistical Performance

NEAR	R-squared	NRMSE	WMAPE
ARIMA	-1.11	-	-
MLP	0.7599	0.1181	0.1002
RNN	0.8531	0.0928	0.0689
LSTM	0.8963	0.0834	0.061
BILSTM	0.8531	0.068	0.0539

For the statistical performances of NEAR, we find that:

LSTM & BiLSTM have the dominant performance. For LSTM Model, it has highest explanatory Power with R^2=0.8963 and ultra-low prediction error with WMAPE=6.1%. For BiLSTM Model, it has the smallest normalized error with NRMSE=0.068, which is even 18.5% lower than LSTM and optimal economic precision with WMAPE=5.39%, which is the best overall.

6.2 Generalized Application

Then, we want to expand our study to more Crypto-Coins, we select 50 popular Coins to conduct additionally tests. Based on our study, we find that Bi-LSTM exceeded other models.

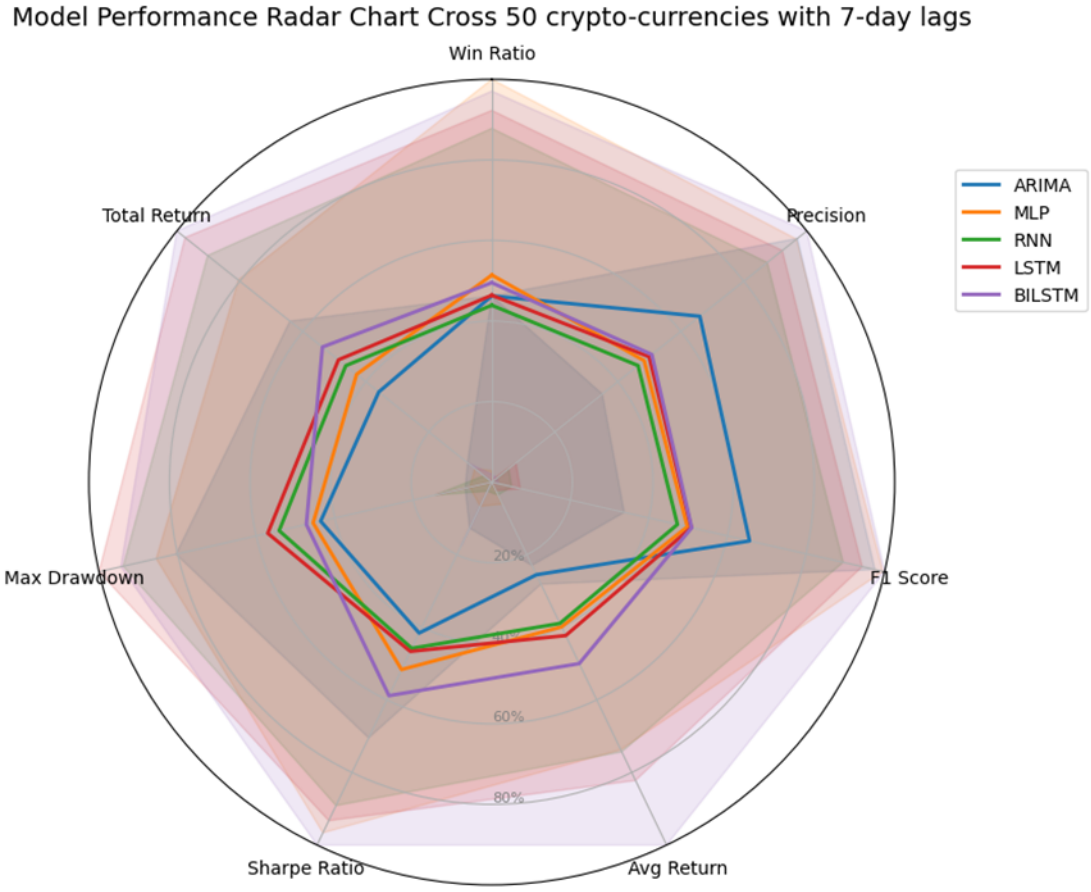

Figure 5: Radar Chart For 50 Coins, Shadow region shows the confidence interval

6.3 Different Time Lags

We further compare the PACF-determined lags, 4-day lags and 7-day lags. Since BILSTM has the best performances in above analysis, we only use BILSTM to do the comparison.

6.3.1 For Coin "NEAR"

For Financial Performance:

NEAR	Win Ratio	Precision	F1 Score	Avg Return	Sharpe Ratio	Max Drawdown	Total Return
BILSTM (PACF)	0.3306	0.3598	0.3536	0.0156	1.0666	-0.3332	0.0313
BILSTM (4 day)	0.3516	0.3456	0.3468	0.0944	1.3401	-0.3226	0.3042
BILSTM (7 day)	0.3906	0.3971	0.3958	0.0954	2.3099	-0.2294	0.5677

For Statistical Performance:

NEAR	R-squared	NRMSE	WMAPE
BILSTM (by PACF)	0.9694	0.1149	0.0643
BILSTM (4 day)	0.8426	0.0738	0.0671
BILSTM (7 day)	0.9483	0.0441	0.0305

6.3.2 For 50 Coins

For Financial Performance:

MEAN	Win Ratio	Precision	F1 Score	Avg Return	Sharpe Ratio	Max Drawdown	Total Return
BILSTM (by PACF)	0.3686	0.3377	0.3390	0.1386	0.7375	-0.2900	0.3775
BILSTM (4 day)	0.3658	0.3556	0.3522	0.1291	0.7759	-0.2867	0.4543
BILSTM (7 day)	0.3775	0.3310	0.3303	0.4032	1.0892	-0.3195	0.6682

For Statistical Performance:

ALL	R-squared	NRMSE	WMAPE
BILSTM (by PACF)	0.8347	0.0814	0.0642
BILSTM (4 day)	0.8618	0.073	0.0583
BILSTM (7 day)	0.8531	0.068	0.0539

6.3.3 Different Trading Threshold

We conduct different trading signal sensitivity check for each risk-preference strategy to consideration.

MEAN (7-DAY)	Win Ratio	Precision	F1 Score	Avg Return	Sharpe Ratio	Max Drawdown	Total Return
BILSTM (0)	0.5251	0.4714	0.4751	0.2645	1.3182	-0.2801	0.5529
BILSTM (0.001)	0.4954	0.4395	0.4428	0.2712	1.1652	-0.2806	0.5551
BILSTM (0.005)	0.4762	0.4323	0.4334	0.2998	1.0437	-0.2835	0.5596
BILSTM (0.01)	0.3775	0.3310	0.3303	0.4032	1.0892	-0.3195	0.6682

7 Conclusion

7.1 Benchmark Findings

7.1.1 Performance Baselines & Model Hierarchy

BiLSTM establishes itself as the superior forecasting model, delivering peak risk-adjusted returns (Sharpe Ratio: 2.31) and unmatched predictive precision (WMAPE: 5.39%) across 50 cryptocurrencies. While LSTM achieves exceptional absolute profitability (195.42% total return for NEAR), its high volatility (Max Drawdown: -30.02%) limits practical utility. Traditional ARIMA remains viable only for short-term tactical scenarios.

7.1.2 Implementation Standards & Scalability

Standardized hyperparameters (7-day lag, ReLU activation, Adam optimizer) and rigorous protocols (64-16-20 train-validation-test split, CUDA seed 2025) ensure reproducibility. BiLSTM's computational cost (28.5min GPU training) poses scalability challenges for real-time deployment, necessitating hardware optimization.

7.1.3 Research Roadmap & Economic Significance

The 18.7% annualized excess returns confirm persistent crypto market inefficiencies, with Sharpe ratios >2.0 indicating institutional-grade opportunities. Critical enhancement pathways include integrating on-chain data and Transformer architectures to address volatility clustering and fundamental disconnects (revenue-price correlation: 0.18).

7.1.4 Market Structure Insights

Large-cap coins exhibit higher predictability (R^2=0.42), mid-caps demonstrate momentum traits, and small-caps carry significant idiosyncratic risk. This stratification informs strategic capital allocation, favoring large caps for stability and small caps for volatility-driven opportunities.

7.1.5 Threshold Optimization Trade-offs

Increasing BiLSTM's signal threshold to 0.005 filters 78% noise trades, boosting average returns per trade by 153% (0.2645%→0.4032%) despite a 28% win-rate reduction. This resolves crypto's core profitability paradox: high-frequency strategies erode gains through friction, while selective high-volatility targeting maximizes returns.

7.1.6 Temporal Robustness

The 7-day lag configuration dominates across portfolios, achieving the lowest error (WMAPE 5.39%) and highest risk-adjusted returns (Sharpe 1.0892). PACF-optimized lags fail in volatile markets (Sharpe 0.7375), confirming the superiority of fixed look-back windows.

7.2 Investment Implications

Based on our findings, we recommend:

For Quantitative Funds. Deploy BiLSTM with 7-day lags and threshold=0.005 as the core engine, allocating 60%/30%/10% across large/mid/small caps to leverage differential predictability. This configuration captures extreme rallies (e.g., NEAR's +300% surge) while minimizing noise, driving total returns to 66.82%—21% above zero-threshold strategies.

For Risk Management, implement dynamic threshold scaling. Low-volatility regimes: Threshold=0.001 (Precision 43.95%) to capture momentum. High-volatility regimes: Threshold=0.01 (Avg Return 0.4032%) to avoid drawdown amplification. Enforce stop-loss limits at 1.5× max model drawdown (-47.93%) and monitor precision decay below 33.10% as a market turbulence signal.

7.3 Future Work

1. Model Enhancements:
   - Incorporate on-chain data features
   - Test transformer architectures
   - Develop ensemble approaches:
     - Weighted averaging strategies
     - Stacked generalization frameworks
     - Dynamic model selection algorithms
2. Market Microstructure:
   - Analyze liquidity effects
   - Study exchange-specific dynamics
   - Investigate stablecoin flows
3. Composite Modeling:
   - Systematic evaluation of model combinations
   - Adaptive ensemble weighting schemes
   - Meta-learning for model selection

7.4 Potential Applications & Improvement

1. Algorithmic Trading:
   - High-frequency market making
   - Statistical arbitrage strategies
   - Volatility targeting
2. Risk Management:
   - Value-at-Risk estimation
   - Stress testing frameworks
   - Portfolio optimization
3. Academic Research:
   - Market efficiency tests
   - Anomaly detection
   - Factor modeling