Jurnal Mandiri IT
Vol. 14 No. 2, October (2025), pp. 159-169

ISSN 2301-8984 (Print), 2809-1884 (Online)
Published by Institute of Computer Science

Reinforcement learning for bitcoin trading: A comparative study of PPO and DQN

Romadhan Edy Prasetyo[1], Sumanto[2], Indra Chaidir[3], Adi Supriyatna[4]
[1,2,3]Informatics Study Program, Faculty of Engineering and Informatics, Universitas Bina Sarana Informatika, Jakarta, Indonesia
[4]Accounting Information Systems Study Program, Faculty of Engineering and Informatics, Universitas Bina Sarana Informatika, Jakarta, Indonesia

A R T I C L E I N F O

Article history:

Received Aug 6, 2025
Revised Aug 12, 2025
Accepted Aug 19, 2025

Keywords:

Bitcoin;
Cryptocurrency Trading;
Deep Q-Network;
Proximal Policy Optimization;
Reinforcement Learning.

ABSTRACT

Bitcoin's high volatility demands automated strategies that adapt to changing market regimes while managing risk. This study compares Proximal Policy Optimization (PPO) and Deep Q-Network (DQN) for Bitcoin trading using hourly BTC/USDT data from 2019 to early 2025. The models are trained to generate buy and sell signals from technical indicators including the Relative Strength Index (RSI), MA20, volatility, Moving Average Convergence Divergence (MACD), volume trend, SMA200, and a weekly trend filter. All features are computed on hourly bars. The evaluation shows that PPO tends to trade more aggressively and delivers higher performance during bullish phases, though with greater risk in unstable markets. By contrast, DQN trades more selectively and maintains better stability in sideways or choppy conditions. These findings support the effectiveness of reinforcement learning for adaptive cryptocurrency trading and highlight complementary strengths between PPO and DQN across market regimes.

Corresponding Author:

Romadhan Edy Prasetyo,
Informatics Study Program,
Faculty of Engineering and Informatics,
Universitas Bina Sarana Informatika,
Jl. Kramat Raya No. 98, Jakarta, 10450, Indonesia
Email: romadhanedy@gmail.com

1. INTRODUCTION

Bitcoin is a highly volatile cryptocurrency, which creates challenges for developing automated trading strategies that remain adaptive. Traditional price-forecasting methods such as ARIMA often struggle to anticipate rapidly changing market dynamics (Indriyanti et al., 2025). As a result, reinforcement learning (RL) has been increasingly adopted to build trading strategies that learn from interaction with the market and enable adaptive buy and sell decisions. Recent studies show that DRL can optimize risk-adjusted objectives directly (e.g., Sharpe) and improve decision quality in volatile markets, supporting its use beyond price prediction (Théate & Ernst, 2021; Zhang et al., 2024).

Baradja and Tjendrowasono (2024) applied Deep Q-Learning to automate foreign-exchange trading and showed that RL can learn useful patterns without explicit supervision. In cryptocurrency spot markets, Huang and Su (2024) developed a Deep Q-Network (DQN) strategy and reported competitive signal accuracy. Complementing this, DADE-DQN introduces dual-action and dual-environment mechanisms to balance exploration and risk, reporting improved trading results versus plain DQN (Y. Huang et al., 2023). In the Indonesian equity market, Saepudin and Rauf (2025) demonstrated the effectiveness of Proximal Policy Optimization (PPO) with risk performance that surpassed Advantage Actor–Critic.

Faturohman and Nugraha (2022) applied deep reinforcement learning to optimize Islamic stock portfolios, showing that RL can adaptively adjust asset allocation to market conditions. Firsov et al. (2023) reported that PPO maintained stable profitability in high-volatility crypto markets and outperformed conventional methods.

Prior work showed that deep-learning models such as artificial neural networks and long short-term memory (LSTM) networks can capture nonlinear price patterns in crypto, but they do not explicitly model trading actions the way RL does (Fegiyanto et al., 2024; Moch Farryz Rizkilloh & Sri Widiyanesti, 2022). More recent RL studies enrich the state with multimodal embeddings (e.g., news and sentiment) and even candlestick-image features, which have yielded stronger PnL and interpretability in crypto trading (Avramelou et al., 2024; Jing & Kang, 2024). Still, cross-market evidence remains mixed, and empirical analyses emphasize careful validation and robustness checks because DRL agents do not always generalize or outperform benchmarks (Kong & So, 2023). Despite these advances, in-depth comparisons between a policy-based algorithm such as PPO and a value-based algorithm such as DQN for trend-conditioned Bitcoin trading remain limited, and quarterly stability/risk evaluations are also scarce. Prior studies include a DRL-based Bitcoin transaction strategy (F. Liu et al., 2021), a multi-level Deep Q-Network for Bitcoin trading (Otabek & Choi, 2024), and deep RL for optimal placement of cryptocurrency limit orders (Schnaubelt, 2022). However, none of these directly compare PPO and DQN under a trend filter or report quarterly analyses.

While PPO and DQN have been applied separately in various financial contexts, direct comparative studies in the cryptocurrency market that integrate a trend filter such as an SMA200 regime condition remain rare. Previous research seldom investigates how a policy-based algorithm like PPO and a value-based algorithm like DQN behave differently when constrained by market trend regimes and evaluated on a quarterly basis. This situation creates a gap in understanding their respective strengths under different market conditions. PPO, with its policy gradient framework, offers rapid adaptation to favorable momentum, whereas DQN, with its value approximation approach, tends to promote more selective and risk-aware trading. These characteristics are particularly relevant for the design of adaptive Bitcoin trading strategies that must operate in environments with high volatility, frequent regime shifts, and the need to filter trades in less favorable trends. Filling this gap can provide practical guidance for selecting algorithms in trend-conditioned cryptocurrency trading.

2. RESEARCH METHOD

Research process and dataset

This study followed a structured pipeline that began with a literature review and continued with dataset acquisition, preprocessing and feature engineering, development of a reinforcement learning trading environment, training of Deep Q-Network and Proximal Policy Optimization models, and performance evaluation. Hourly BTC/USDT spot prices were retrieved from the Binance public REST API for January 2019 through March 2025. Each record included timestamp, open, high, low, close, and volume. The dataset contained more than 50,000 observations. Data were stored in tabular form and split chronologically into a training set for 2019 to 2022 and an out-of-sample test set for 2023 to March 2025 to assess generalization under recent market conditions.

Preprocessing and technical feature extraction

Preprocessing checked for missing values, duplicated timestamps, and timing anomalies. Gaps caused by API interruptions were handled through removal or imputation as appropriate to maintain continuity. Timestamps were converted to a uniform timezone, and all indicators were computed on synchronized hourly bars.

The technical features used as model inputs were the relative strength index (RSI, period 14), a 20-period moving average (MA20), price volatility computed over the last 20 hours, moving average convergence divergence (MACD with 12/26 EMAs and a 9-period signal), a simple volume-trend measure based on first differences in volume, a 200-period simple moving average (SMA200) that later served as a regime indicator, and a weekly trend filter. The weekly trend was derived from resampled weekly closes and then forward-aligned to the hourly timeline so that every timestamp carried a consistent weekly signal. To prevent price levels from dominating learning, the close price was standardized by Z-score over the historical window, while the remaining indicators were kept in their native scales to preserve their interpretability.

Reinforcement learning environment

The trading environment followed the Gymnasium interface, as described by Towers et al. (2024), and defined a discrete action set of three choices (hold, buy, and sell). The policy followed an all-in/all-out rule, whereby at each hour the agent was either entirely in cash or entirely in BTC. Actions were executed at the hourly close. The state vector at each step contained ten elements, namely standardized close, RSI, MA20, volatility, MACD, volume trend, cash balance, BTC balance, a binary regime flag based on SMA200 (1 bullish when price > SMA200, 0 bearish otherwise), and the weekly trend score in the range −1 to +1. This design follows the safety-layer idea by adding rule-based guardrails to reduce tail risks while the agent optimizes returns (Kochliaridis et al., 2023).

The reward combined profitability, realized gains on sells, drawdown control, and small regime-aware signals (Y. Huang et al., 2024). Transaction costs were set to zero in simulation to isolate algorithmic behavior from fee effects. The reward used during training was

$$r_t = (profit_t + 0.5 \times sell\_profit_t + fee\_penalty_t - 0.3 \times drawdown_t) \times regime\_factor_t + 0.1 \times weekly\_trend_t$$

Definitions

- r_t denotes the reward at time *t*.
- $profit_t$ denotes the step change in net worth from *t−1* to *t*.
- $sell\_profit_t$ denotes the realized profit from the most recent sell at *t* (zero when no sell occurs).
- $fee\_penalty_t$ denotes the transaction fee at *t* and is set to zero in this study.
- $drawdown_t$ denotes the drawdown at *t*, computed as the percentage drop from the running peak of net worth.
- $regime\_factor_t$ denotes the trend-condition factor derived from SMA200 and equals 1.0 when price > SMA200 (bullish) and 0.8 when price < SMA200 (bearish).
- $weekly\_trend_t$ denotes the weekly trend score in the range −1 to +1, providing a small bonus or penalty aligned with higher-timeframe direction.

An early stopping rule was applied at the episode level. If net worth fell to 50 percent of initial capital, the episode ended immediately. This cutoff prevented long trajectories with extreme losses and reduced exploration of harmful policies. Evaluation proceeded strictly chronologically without lookahead until the data ended or the early stopping criterion was met.

Algorithms and Implementation

Two algorithms were evaluated. DQN represented a value-based approach that learned an action-value function with experience replay and a target network (Mnih et al., 2015). PPO represented a policy-based approach that updated the policy via a clipped surrogate objective for stable improvement (Schulman et al., 2017). Both were implemented with Stable-Baselines3 to ensure reproducibility and standardized utilities (Raffin et al., 2021).

Training Procedure

Models were trained on hourly BTC/USDT data (2019–2022) in a Gymnasium environment using Stable-Baselines3, with a fixed random seed (42) and 300,000 timesteps per run. The environment used DummyVecEnv and Monitor for training and performance tracking. PPO (MlpPolicy, on-policy) was tuned over learning rates {2×10^{-4}, 1×10^{-4}} and n_steps {1024, 2048}, with batch_size = 64, ent_coef = 1×10^{-4}, γ = 0.98, gae_λ = 0.92, clip_range = 0.2, and target_kl = 0.01 for KL-based early stopping. DQN (MlpPolicy Q-network, off-policy) was tuned over learning rates {1×10^{-4}, 5×10^{-5}} and batch sizes {64, 128}, with buffer_size = 100,000, learning_starts = 5,000, train_freq = 4, γ = 0.99, τ = 0.05, target_update_interval = 1,000, exploration_fraction = 0.3, exploration_final_ε = 0.05, and gradient_steps = 1. A RewardTrackingCallback recorded rewards, and performance was evaluated using the final reward (mean reward of the last 100 timesteps). Models were saved with hyperparameter-based names, and the best configuration for each algorithm was selected based on the highest final reward. Our PPO setup, including KL-based early stopping, and Gymnasium-style market environment follow FinRL-Meta's DRL-trading framework (X.-Y. Liu et al., 2022). DQN training ran for the full 300,000 timesteps without early stopping, consistent with DRL-trading practices (Théate & Ernst, 2021).

Data processing and analysis

Post-training analysis focused on the out-of-sample window. The aim was to test the strategies learned by PPO and DQN and to compare them with a passive benchmark. The analysis sequence comprised signal generation on test data, portfolio backtesting, metric computation, and comparative interpretation.

Signal generation and backtesting

The best PPO and DQN models were run chronologically from January 2023 to March 2025. The agents read hourly indicators at each timestamp and produced one of three actions. Forward testing was strictly chronological so the models never accessed future data. To keep results reproducible, a fixed random state was set before evaluation.

Backtesting used an initial portfolio of USD 10,000 in cash with no BTC. Buy invested the full cash balance into BTC, sell liquidated the full BTC position into cash, and hold left the position unchanged. Evaluation was conducted per calendar quarter. At the start of each quarter the portfolio was reset to USD 10,000 so that quarterly comparisons remained independent. As a benchmark, a passive buy-and-hold strategy bought BTC worth USD 10,000 at the start of the period and held it to quarter end. For each quarter the procedure produced the ending portfolio values and the corresponding time series for risk analysis.

Evaluation protocol and metrics

We evaluated the policies exclusively on the out-of-sample window and summarized performance by calendar quarter to capture regime shifts without mixing periods. For each quarter and for each strategy (PPO, DQN), we computed: a) Return on Investment: Cumulative return per quarter, with portfolio reset to USD 10,000; b) Sharpe Ratio: Mean hourly simple return divided by its standard deviation, annualized (24×365=8,760 periods); c) Sortino Ratio: Mean hourly simple return divided by downside deviation, annualized; d) Realized Volatility: Standard deviation of hourly simple returns, annualized; d) Maximum Drawdown: Largest peak-to-trough decline per quarter; e) Win Rate: Fraction of steps with increasing net worth; f) Trade Frequency: Number of buy/sell trades.

Backtests used a fixed seed (42) and set_rng_state_to_step (seed=42, step=5) for reproducibility, with USD 10,000 initial capital and no further learning. Sharpe and Sortino assumed a zero risk-free rate). Hourly simple returns may cause ROI and Sortino sign differences in edge cases with high volatility. Signal patterns were analyzed across bullish/bearish regimes.

3. RESULTS AND DISCUSSIONS

Training results

Models were trained on Bitcoin data from 2019 to 2022 using PPO and DQN. The goal was to identify hyperparameter settings that produced the highest final reward. After tuning, results from all configurations were recorded and compared.

Table 1. PPO training results

model_path	learning_rate	n_steps	final_reward
models/ppo_lr2e-04_n2048.zip	0.0002	2048	-0.113122314
models/ppo_lr2e-04_n1024.zip	0.0002	1024	-0.13879122
models/ppo_lr1e-04_n2048.zip	**0.0001**	**2048**	**-0.035938982**
models/ppo_lr1e-04_n1024.zip	0.0001	1024	-0.06595223

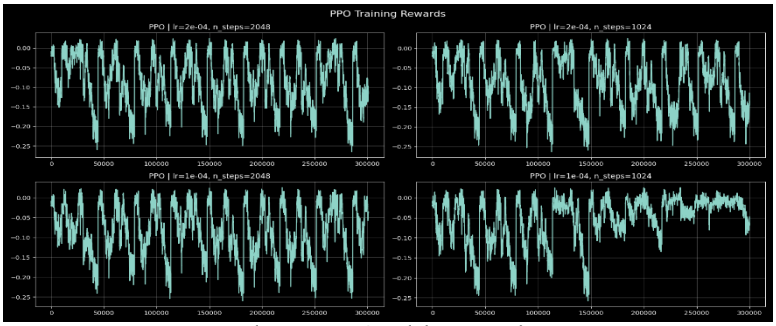

Figure 1. PPO training rewards

The best PPO setting used a learning rate of 0.0001 with 2,048 steps per update, yielding a final reward of −0.0359. Other settings produced lower final rewards in absolute terms.

Table 2. DQN training results

model_path	learning_rate	batch_size	final_reward
models/dqn_lr1e-04_batch128.zip	0.0001	128	-0.008028825
models/dqn_lr1e-04_batch64.zip	0.0001	64	-0.072349764
models/dqn_lr5e-05_batch128.zip	0.00005	128	-0.09587393
models/dqn_lr5e-05_batch64.zip	0.00005	64	-0.16240415

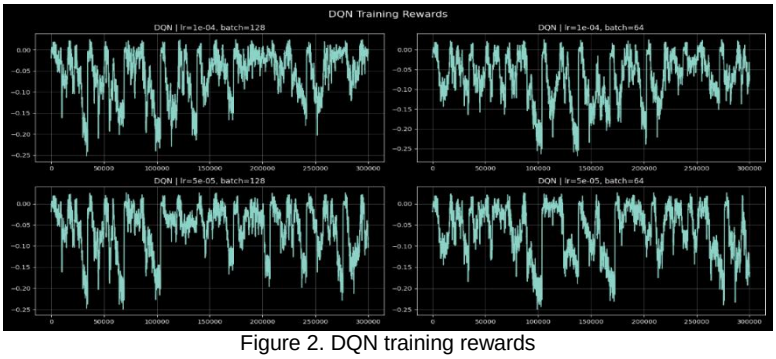

Figure 2. DQN training rewards

The best DQN setting used a learning rate of 0.0001 with a batch size of 128, yielding a final reward of −0.0080. Smaller learning rates or batch sizes did not improve the outcome. Final rewards are negative by design, making values nearer zero preferable. Training curves support this and reveal hyperparameter sensitivity, while selection used final reward with the same number of timesteps and the same random seed (42), followed by out-of-sample validation on 2023 to 2025 data.

Trading signals

Forward testing produced buy and sell signals on the test window. Table 3 summarizes signal counts by quarter. PPO generally generated more signals than DQN in most quarters, which reflects a more active response to price movements. An exception occurred in 2023Q1 when DQN issued more trades. In contrast, DQN was typically more selective.

Table 3. Quarterly trading signals summary

Model	Year-Quarter	Buy Signals	Sell Signals	Total Signals
PPO	2023Q1	67	66	133
PPO	2023Q2	44	43	87
PPO	2023Q3	44	43	87
PPO	2023Q4	34	33	67
PPO	2024Q1	43	42	85
PPO	2024Q2	45	45	90
PPO	2024Q3	42	41	83
PPO	2024Q4	39	38	77
PPO	2025Q1	48	47	95
DQN	2023Q1	88	87	175
DQN	2023Q2	27	26	53
DQN	2023Q3	21	21	42
DQN	2023Q4	16	16	32
DQN	2024Q1	18	17	35
DQN	2024Q2	18	18	36
DQN	2024Q3	22	21	43
DQN	2024Q4	18	18	36
DQN	2025Q1	20	19	39

PPO tended to be more aggressive than DQN in generating trading signals. PPO produced a high number of signals in most quarters, with dense distributions across time. In contrast, DQN was more selective, issuing fewer signals that tended to appear only under specific market conditions.

Figure 3, Figure 4, and Figure 5 visualize PPO buy and sell signals on the 2023, 2024, and 2025 test data, respectively. Figure 6, Figure 7, and Figure 8 visualize the corresponding DQN signals for the same years.

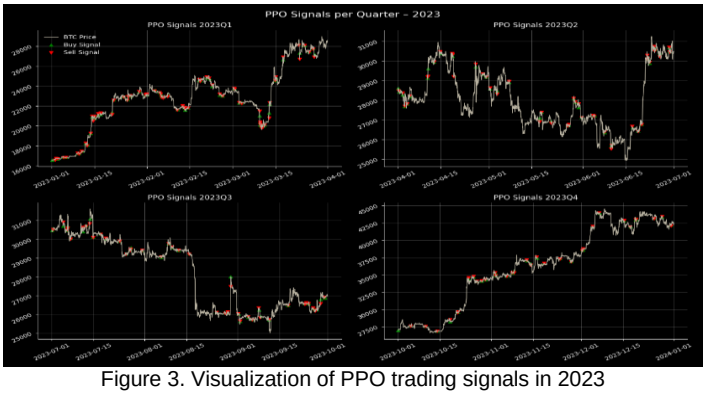

Figure 3. Visualization of PPO trading signals in 2023

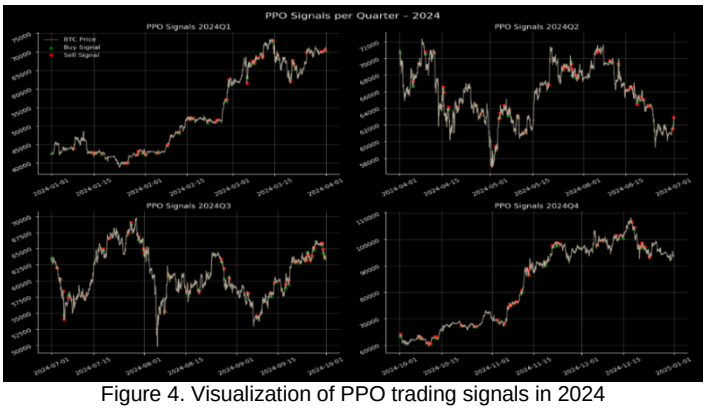

Figure 4. Visualization of PPO trading signals in 2024

Figure 5. Visualization of PPO trading signals in 2025

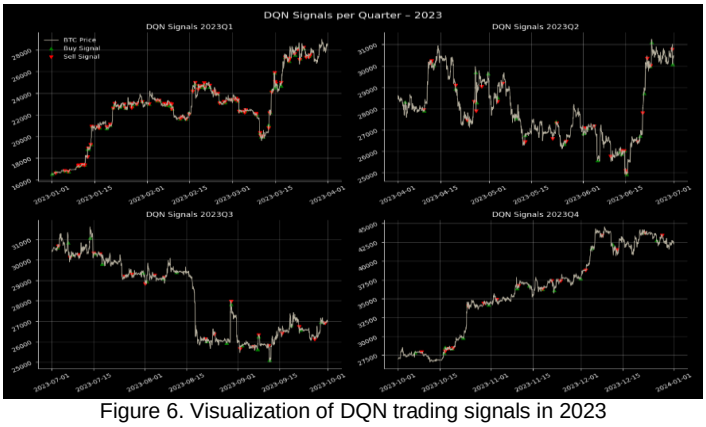

Figure 6. Visualization of DQN trading signals in 2023

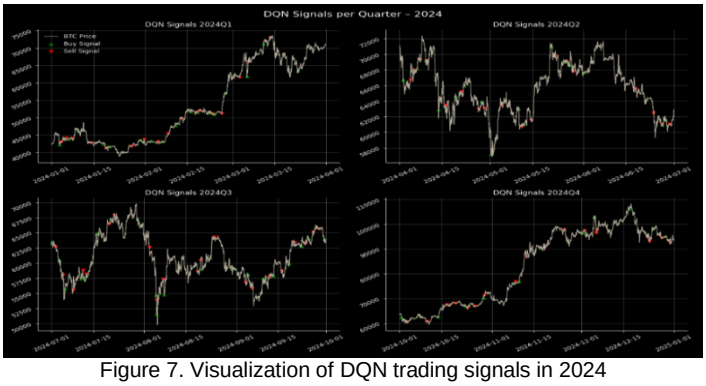

Figure 7. Visualization of DQN trading signals in 2024

Figure 8. Visualization of DQN trading signals in 2025

Portfolio Growth

Figures 9 to 11 plot the equity curves with an initial capital of USD 10,000 at the start of each quarter. The portfolio is reset at each quarter boundary so comparisons remain independent. Execution followed the same rules as in the evaluation, with actions taken at the hourly close, long-only unit-sized positions, zero transaction costs, and an early stop when net worth fell to 50 percent of initial capital. If the early stop triggered inside a quarter, the curve ended at that time and remained flat for the rest of the quarter.

PPO showed sharper swings, with strong run-ups during clear uptrends and steeper drawdowns during corrections. This was visible in the early part of 2023 and in 2024 when trends strengthened, as well as in early 2025 when pullbacks deepened. DQN produced a smoother and more stable trajectory with more moderate gains. The difference reflects trading style, with PPO reacting quickly and compounding momentum when the market is favorable, whereas DQN trades less often and preserves capital better in choppy or uncertain conditions.

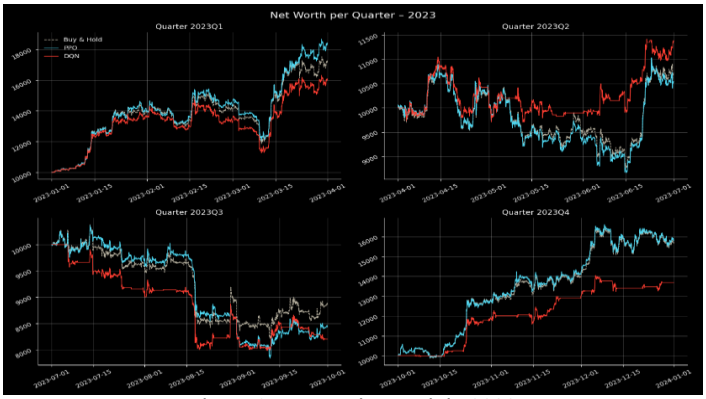

Figure 9. Net worth growth in 2023

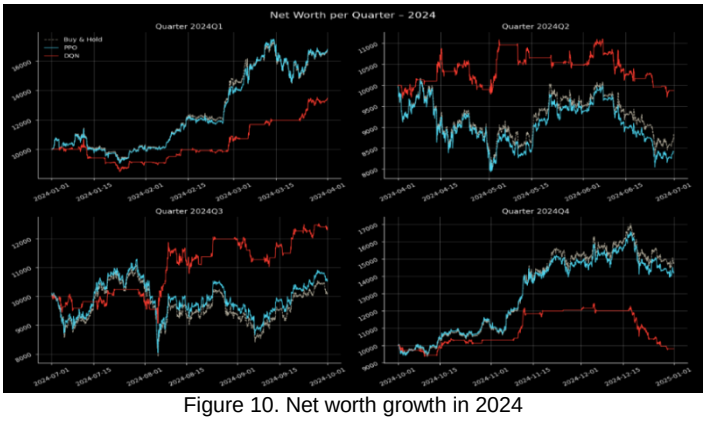

Figure 10. Net worth growth in 2024

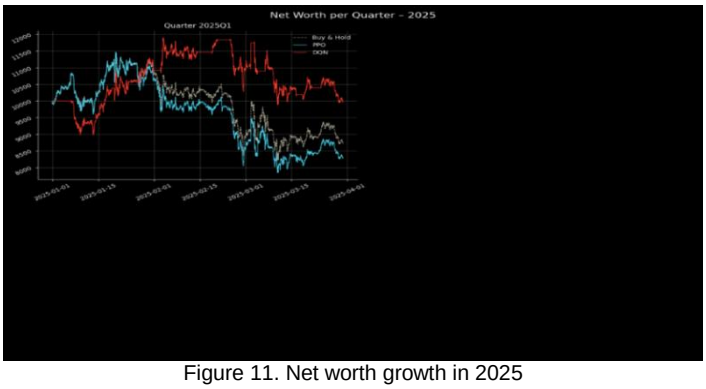

Figure 11. Net worth growth in 2025

Quantitative evaluation

Performance was assessed each quarter using return, Sharpe ratio, Sortino ratio, volatility, maximum drawdown, win rate, and trade frequency. Table 4 and Table 5 report results for PPO and DQN from 2023Q1 to 2025Q1.

Table 4. Quarterly evaluation metrics for PPO

Year-Quarter	Return	Sharpe Ratio	Sortino Ratio	Volatility	Max Drawdown	Win Rate	Trading Freq
2023Q1	83.48%	5.10	6.80	50.72%	21.62%	50.70%	133
2023Q2	5.41%	0.66	0.76	41.47%	20.41%	49.11%	87
2023Q3	-15.66%	-2.14	-2.34	29.37%	24.30%	49.30%	87
2023Q4	57.93%	4.69	6.65	40.33%	8.60%	51.20%	67
2024Q1	68.24%	4.08	5.44	54.57%	20.56%	50.48%	85
2024Q2	-15.90%	-1.18	-1.41	47.79%	21.68%	49.02%	90
2024Q3	5.23%	0.56	0.74	54.71%	28.79%	49.12%	83
2024Q4	41.97%	3.03	4.00	50.17%	16.01%	51.34%	77
2025Q1	-17.17%	-1.06	-1.31	55.32%	31.57%	49.04%	95

PPO delivered higher returns in quarters with strong upward trends, such as 2023Q1 (83.48%) and 2024Q1 (68.24%). These quarters also showed high Sharpe and Sortino ratios, which indicates efficient use of risk when the market trended higher. At the same time, PPO recorded higher volatility and larger drawdowns alongside consistently high trade counts. The combination implies greater risk exposure during unstable conditions, as reflected by the negative return in 2025Q1 (−17.17%) with a drawdown of 31.57%.

Table 5. Quarterly evaluation metrics for DQN

Year-Quarter	Return	Sharpe Ratio	Sortino Ratio	Volatility	Max Drawdown	Win Rate	Trading Freq
2023Q1	60.28%	3.91	4.94	52.33%	23.53%	49.86%	175
2023Q2	13.76%	1.77	1.84	32.22%	12.10%	34.77%	53
2023Q3	-17.94%	-2.74	-2.43	27.29%	20.09%	28.73%	42
2023Q4	36.78%	4.75	4.42	26.95%	5.81%	19.35%	32
2024Q1	34.88%	3.52	3.55	36.00%	20.80%	25.47%	35
2024Q2	-1.27%	-0.01	-0.01	31.12%	12.56%	25.24%	36
2024Q3	23.05%	2.21	2.22	40.93%	11.59%	25.24%	43
2024Q4	-1.91%	-0.10	-0.08	30.14%	21.57%	16.09%	36
2025Q1	-0.27%	0.21	0.24	47.42%	16.90%	34.38%	39

DQN followed a more conservative pattern. While returns were often lower than PPO during bullish phases, DQN tended to control risk more consistently. In quarters such as 2023Q4 and 2024Q4, DQN posted materially lower maximum drawdowns than PPO. Lower trade frequency supports the view that DQN acted more selectively. Even in bullish quarters like 2023Q1 and 2024Q1, DQN's returns remained competitive at 60.28% and 34.88%.

Sortino ratios varied substantially across quarters for both strategies. Negative or near-zero values in some periods signal that downside risk remained material. For example, in 2023Q3 and 2025Q1, both strategies recorded negative Sortino ratios, indicating that gains were not sufficient to offset downside variability in those conditions.

Two edge cases help interpret the quarterly results. A quarter can end with a gain yet show a low or negative Sortino ratio when many negative periods or a few large losses depress the average periodic return while the downside deviation remains high. The reverse can also occur, as in DQN 2025Q1 where the quarter ends with a small loss while the Sortino ratio is slightly positive. Quarterly ROI is a simple change in net worth. The Sortino ratio is produced from hourly returns and then annualized using 24 times 365 periods. Differences in aggregation and the distribution of downside periods within the quarter can therefore yield the sign patterns observed.

Discussion of findings

The results highlight distinct behavioral profiles. Similar patterns are noted elsewhere: multimodal/ensemble RL often amplifies momentum capture in favorable regimes but needs safeguards to limit drawdowns during corrections (Avramelou et al., 2024; Jing & Kang, 2024; Kochliaridis et al., 2023). PPO is advantageous when a clear uptrend is present because it reacts quickly and compounds gains, but it is more exposed to false signals and deeper drawdowns in choppy or declining markets. DQN is better suited to uncertain or sideways markets because it trades less, controls drawdowns more effectively, and preserves stability at the cost of occasionally missing momentum in strong rallies. This aligns with prior evidence that purely predictive approaches such as ARIMA may struggle to anticipate rapid regime shifts in crypto markets, while reinforcement learning can adapt decisions to feedback from the environment (Indriyanti et al., 2025).

Overall, PPO and DQN offer complementary strengths. Emerging pro-trader RL frameworks explicitly encode professional trading patterns and risk discipline, pointing to a practical path for combining PPO's trend responsiveness with DQN's stability (Jeong & Gu, 2024). PPO fits conditions with strong, sustained trends, whereas DQN is more effective during corrections and uncertainty.

4. CONCLUSION

This study compared Proximal Policy Optimization (PPO) and Deep Q-Network (DQN) for Bitcoin trading using historical data and standard technical indicators, and evaluated the models out-of-sample from 2023 to March 2025 with a bullish market filter based on the SMA200. The results

showed clear behavioral differences. PPO generated a higher frequency of signals and captured larger cumulative gains during strong uptrends, yet it also faced higher volatility and deeper maximum drawdowns when markets consolidated or corrected. DQN traded more selectively and produced steadier portfolios with lower drawdowns in several quarters, although its returns were typically more moderate than PPO.

These findings translate into practical recommendations for other liquid crypto markets. First, use a simple regime filter such as a long horizon moving average or an equivalent trend indicator to gate participation and to reduce exposure when the market is unfavorable. Second, select the algorithm by regime. PPO is better suited when momentum is strong and persistent, while DQN is preferable in uncertain or sideways conditions that reward selective entry. Third, retune hyperparameters per asset and revalidate by quarter or by regime to account for differences in liquidity and volatility. Fourth, incorporate realistic frictions and execution constraints such as transaction costs, slippage assumptions, turnover caps, and cooldown rules, since these can materially change the net outcome.

The results also imply concrete design choices for long term risk management in algorithmic trading systems. A simple and robust approach is to allocate capital dynamically between the two agents based on a regime score so that a conservative DQN allocation is maintained in ambiguous markets while a PPO allocation is activated when trend conviction is high. Risk overlays should be applied at the portfolio layer. Examples include volatility targeting to stabilize risk, maximum position size to prevent concentration, a drawdown based stop trading rule to cap tail losses, and limits on trade frequency to control costs. Periodic retraining and drift monitoring are recommended so that policies remain aligned with recent market structure. Evaluation should remain strictly out of sample and should include rolling windows to test stability over time.

In practical terms, PPO and DQN are complementary. PPO is an effective momentum capture component in favorable regimes, while DQN serves as a stability anchor when conditions are noisy or directionless. Combining a regime filter, a switching rule between agents, and portfolio level risk controls provides a viable blueprint for extending this study to other crypto assets with varying microstructure and volatility.

ACKNOWLEDGEMENTS

The authors thank the Informatics Study Program and the Faculty of Engineering and Informatics, Universitas Bina Sarana Informatika, for academic support and access to computing resources. We are grateful to our academic advisors and colleagues for constructive feedback throughout the study. We also acknowledge Binance for the public BTC/USDT market data and the maintainers of Gymnasium and Stable-Baselines3, whose open-source tools facilitated the experiments. This research received no specific grant from any funding agency, commercial or not-for-profit sectors. Any remaining errors are our own.

REFERENCES

Avramelou, L., Nousi, P., Passalis, N., & Tefas, A. (2024). Deep reinforcement learning for financial trading using multi-modal features. *Expert Systems with Applications*, *238*, 121849. https://doi.org/10.1016/j.eswa.2023.121849

Baradja, A., & Tjendrowasono, T. I. (2024). Pengaplikasian Deep Reinforcement Q-Learning Untuk Prediksi Perdagangan Valas Otomatis. *Jurnal Rekayasa Sistem Informasi Dan Teknologi*, *1*(3), 190–198. https://doi.org/10.59407/jrsit.v1i3.519

Faturohman, T., & Nugraha, T. (2022). ISLAMIC STOCK PORTFOLIO OPTIMIZATION USING DEEP REINFORCEMENT LEARNING. *Journal of Islamic Monetary Economics and Finance*, *8*(2), 181–200. https://doi.org/10.21098/jimf.v8i2.1430

Fegiyanto, R., Hermawan, A., & Ardiani, F. (2024). Prediksi Harga Crypto dengan Algoritma Jaringan Saraf Tiruan. *Jurnal Indonesia : Manajemen Informatika Dan Komunikasi*, *5*(3), 2265–2275. https://doi.org/10.35870/jimik.v5i3.728

Firsov, D. V., Silvestrov, S. N., Kuznetsov, N. V., Zolotarev, E. V., & Pobyvaev, S. A. (2023). Using PPO Models to Predict the Value of the BNB Cryptocurrency. *Emerging Science Journal*, *7*(4), 1206–1214. https://doi.org/10.28991/ESJ-2023-07-04-012

Huang, C. S. J., & Su, Y.-S. (2024). Trading Strategy of the Cryptocurrency Market Based on Deep Q-Learning Agents. *Applied Artificial Intelligence*, *38*(1). https://doi.org/10.1080/08839514.2024.2381165

Huang, Y., Lu, X., Zhou, C., & Song, Y. (2023). DADE-DQN: Dual Action and Dual Environment Deep Q-Network for Enhancing Stock Trading Strategy. *Mathematics*, *11*(17), 3626. https://doi.org/10.3390/math11173626

Huang, Y., Zhou, C., Zhang, L., & Lu, X. (2024). A Self-Rewarding Mechanism in Deep Reinforcement Learning for Trading Strategy Optimization. *Mathematics*, *12*(24), 4020. https://doi.org/10.3390/math12244020

Indriyanti, I., Ichsan, N., Fatah, H., Wahyuni, T., & Ermawati, E. (2025). Prediksi Jangka Pendek Harga Bitcoin Dengan Metode Arima. *INTECOMS: Journal of Information Technology and Computer Science*, *8*(1), 163–167. https://doi.org/10.31539/intecoms.v8i1.14446

Jeong, D. W., & Gu, Y. H. (2024). Pro Trader RL: Reinforcement learning framework for generating trading knowledge by mimicking the decision-making patterns of professional traders. *Expert Systems with Applications*, *254*, 124465. https://doi.org/10.1016/j.eswa.2024.124465

Jing, L., & Kang, Y. (2024). Automated cryptocurrency trading approach using ensemble deep reinforcement learning: Learn to understand candlesticks. *Expert Systems with Applications*, *237*, 121373. https://doi.org/10.1016/j.eswa.2023.121373

Kochliaridis, V., Kouloumpris, E., & Vlahavas, I. (2023). Combining deep reinforcement learning with technical analysis and trend monitoring on cryptocurrency markets. *Neural Computing and Applications*, *35*(29), 21445–21462. https://doi.org/10.1007/s00521-023-08516-x

Kong, M., & So, J. (2023). Empirical Analysis of Automated Stock Trading Using Deep Reinforcement Learning. *Applied Sciences*, *13*(1), 633. https://doi.org/10.3390/app13010633

Liu, F., Li, Y., Li, B., Li, J., & Xie, H. (2021). Bitcoin transaction strategy construction based on deep reinforcement learning. *Applied Soft Computing*, *113*, 107952. https://doi.org/10.1016/j.asoc.2021.107952

Liu, X.-Y., Xia, Z., Rui, J., Gao, J., Yang, H., Zhu, M., Wang, C., Wang, Z., & Guo, J. (2022). FinRL-Meta: Market Environments and Benchmarks for Data-Driven Financial Reinforcement Learning. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, & A. Oh (Eds.), *Advances in Neural Information Processing Systems* (Vol. 35, pp. 1835–1849). Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2022/file/0bf54b80686d2c4dc0808c2e98d430f7-Paper-Datasets_and_Benchmarks.pdf

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., Graves, A., Riedmiller, M., Fidjeland, A. K., Ostrovski, G., Petersen, S., Beattie, C., Sadik, A., Antonoglou, I., King, H., Kumaran, D., Wierstra, D., Legg, S., & Hassabis, D. (2015). Human-level control through deep reinforcement learning. *Nature*, *518*(7540), 529–533. https://doi.org/10.1038/nature14236

Moch Farryz Rizkilloh, & Sri Widiyanesti. (2022). Prediksi Harga Cryptocurrency Menggunakan Algoritma Long Short Term Memory (LSTM). *Jurnal RESTI (Rekayasa Sistem Dan Teknologi Informasi)*, *6*(1), 25–31. https://doi.org/10.29207/resti.v6i1.3630

Otabek, S., & Choi, J. (2024). Multi-level deep Q-networks for Bitcoin trading strategies. *Scientific Reports*, *14*(1), 771. https://doi.org/10.1038/s41598-024-51408-w

Raffin, A., Hill, A., Gleave, A., Kanervisto, A., Ernestus, M., & Dormann, N. (2021). Stable-Baselines3: Reliable Reinforcement Learning Implementations. *Journal of Machine Learning Research*, *22*(268), 1–8. https://jmlr.org/papers/v22/20-1364.html

Saepudin, D., & Rauf, K. (2025). Application of Deep Reinforcement Learning for Stock Trading on The Indonesia Stock Exchange. *Jurnal Nasional Pendidikan Teknik Informatika (JANAPATI)*, *14*(1), 144–157. https://doi.org/10.23887/janapati.v14i1.83775

Schnaubelt, M. (2022). Deep reinforcement learning for the optimal placement of cryptocurrency limit orders. *European Journal of Operational Research*, *296*(3), 993–1006. https://doi.org/10.1016/j.ejor.2021.04.050

Schulman, J., Wolski, F., Dhariwal, P., Radford, A., & Klimov, O. (2017). *Proximal Policy Optimization Algorithms*.

Théate, T., & Ernst, D. (2021). An application of deep reinforcement learning to algorithmic trading. *Expert Systems with Applications*, *173*, 114632. https://doi.org/10.1016/j.eswa.2021.114632

Towers, M., Kwiatkowski, A., Terry, J., Balis, J. U., De Cola, G., Deleu, T., Goulão, M., Kallinteris, A., Krimmel, M., KG, A., Perez-Vicente, R., Pierré, A., Schulhoff, S., Tai, J. J., Tan, H., & Younis, O. G. (2024). *Gymnasium: A Standard Interface for Reinforcement Learning Environments*.

Zhang, J., Cai, K., & Wen, J. (2024). A survey of deep learning applications in cryptocurrency. *IScience*, *27*(1), 108509. https://doi.org/10.1016/j.isci.2023.108509