Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Optimization of dynamic crypto portfolios via deep reinforcement learning against supervised learning under transaction costs

Kamal El Kehal

kamal.elkehal@usmba.ac.ma

Sidi Mohamed Ben Abdellah University

Drissia Ennagoura

Sidi Mohamed Ben Abdellah University

Othmane Bakkas

Sidi Mohamed Ben Abdellah University

Mohammed El mahjouby

Abdelmalek Essaadi University

Badre Bossoufi

Sidi Mohamed Ben Abdellah University

Khalid El Fahssi

Sidi Mohamed Ben Abdellah University

Mohamed El Far

Sidi Mohamed Ben Abdellah University

Mohamed Taj Bennani

Sidi Mohamed Ben Abdellah University

Research Article

Keywords: Artificial Intelligence, Cryptocurrency, LSTM, Portfolio Optimization, Random Forest, Deep Reinforcement Learning, Market Alpha

Posted Date: November 4th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11167973/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Optimization of dynamic crypto portfolios via deep reinforcement learning against supervised learning under transaction costs

Kamal El Kehal[1*], Drissia Ennagoura[1,2] ,Othmane Bakkas[1], Mohamed El Mahjouby[3], Badre Bossoufi[4], Khalid El Fahssi[1], Mohamed El Far[1], Mohamed Taj Bennani[1]

[1] Laboratory of Computer Science, Innovation, and Artificial Intelligence (L3IA), Faculty of Science Dhar El Mahraz, Sidi Mohamed Ben Abdellah University, Fez, Morocco

[2] Smartilab Laboratory, Moroccan School of Engineering Sciences (EMSI), Rabat, Morocco

[3] Department of Computer Science, Laboratory of Information Systems and Software Engineering (SIGL), Abdelmalek Essaadi University, Larache, Morocco

[4] Laboratory of Engineering Modelling and Systems Analysis (LIMAS), Sidi Mohamed Ben Abdellah University, Fez, Morocco

** Corresponding author: Kamal EL Kehal, kamal.elkehal@usmba.ac.ma*

Abstract

The management of a cryptocurrency portfolio is difficult since it involves the high volatility and non-stationarity of the cryptocurrency market. Therefore, financial management strategies cannot be directly applied. In this paper, a deep reinforcement learning optimization technique (DRL-PPO) is discussed and compared with supervised predictive models like LSTM, Random Forest, and XGBoost in the task of optimization of a portfolio with ten different cryptocurrencies, taking into account the transaction fee of 0.1%. The results have shown that DRL-PPO outperforms other methods in this case, providing an annualized return of 37.22% and Sharpe ratio of 0.87. It is worth mentioning that the DRL-PPO technique is market neutral (beta = –0.08) and alpha equals 46.41%. Meanwhile, the disadvantages of supervised predictive models are revealed in periods of reversal of the trend; LSTM provides 6.75%, while Random Forest provides 23.54%, being subject to fluctuations in the market (beta approximately equal to 1). Finally, XGBoost provides a negative number (-1.48%) which does not even cover the transaction fees. Summarizing, the success of DRL-PPO is not based on directional accuracy but rather on the ability to optimize trading efficiency and manage risks.

Keywords: Artificial Intelligence; Cryptocurrency; LSTM; Portfolio Optimization; Random Forest; Deep Reinforcement Learning, Market Alpha,

3.1 Introduction

Decentralized finance and cryptocurrency investments exhibit annual volatilities exceeding 80%, have high-dimensional non-linear relations, and are characterized by highly non-stationary return distributions with sudden regime changes [1], [2]. Within this stochastic framework, active risk mitigation and dynamic capital preservation form key problems in the field of quantitative portfolio management [3], [4]. For several decades, modern portfolio optimization was based on the mean-variance approach developed by Markowitz [1], which aimed at balancing expected return and variance of portfolio returns [2]. However, classical econometric approaches relied on several assumptions: stationary prices, constant asset correlations, and multivariate normally distributed returns [1],[2]. Empirical literature suggests that these parametric assumptions break down during the period of liquidity shocks and crises in digital assets markets, making static mean-variance optimization inefficient even in comparison with simple diversification rules [2], [3]. In order to tackle these structural problems, the field of computational finance turned towards Machine Learning (ML) and Deep Learning (DL) algorithms [4]. Recurrent neural models like Long Short-Term Memory (LSTM) networks are used to capture temporal memory and momentum properties of noisy financial time-series [5]. Additionally, tree-based ensemble models like Random Forest (RF) and extreme Gradient Boosting (XGBoost) proved useful in isolating non-linear interactions while dealing with high-frequency noise [6],[7]. While these advancements in methodological techniques still face one key issue known as the implementation gap [8]. Supervised architectures work in a disjoint "predict-then-optimize" framework: models are trained for minimizing forecast loss (Mean Squared Error) regardless of market microstructure [8]. As a result, even tiny directional forecasting errors lead to excessive rebalancing signals [8], [9]. Due to realistic execution fees between 0.1% to 0.25% per trade, the constant turnover of assets leads to a high transaction cost drag, turning the theoretical gross alpha to flat or negative net performance [8],[9]. Additionally, supervised models have high market directionality ($Beta \sim 1$), thus failing to de-risk portfolios and protect capital during prolonged bears [3], [9]. In contrast, Deep Reinforcement Learning (DRL) methodology tackles these issues through reformulation of the problem as a continuous Markov Decision Process (MDP) [10]. Rather than predicting asset prices, a DRL agent directly learns to map the states of the observed environment to the actions of portfolio allocation through trial and error ($\pi : S \rightarrow A$)[10]. Among various formulations of policy gradient, Proximal Policy Optimization (PPO) provides robust convergence [11], [12]. Through introducing a clipped surrogate objective function, PPO constraints the size of the policy update step, avoiding a sudden collapse of the performance in the turbulent regimes [11], [12]. Modern literature emphasizes the crucial importance of reward engineering in financial DRL [13]. While the unconstrained formulations of the reward in terms of profit/loss typically produce highly volatile portfolio allocations, risk-averse friction-aware reward formulations increase Sharpe ratios and limit drawdowns [13], [14]. The explicit penalization of the L_1-norm of portfolio weights adjusts the policy gradients to take execution frictions into account, forcing the agent to rebalance the portfolio only if the expected marginal gain exceeds the transaction cost [14], [15]. In order to deal with both empirical and methodological challenges, this paper introduces an end-to-end friction-aware DRL-PPO algorithm applied to top ten cryptocurrencies by market capitalization [10]. Contributions of this work are: Friction-Aware Portfolio Allocation: Directly embedding a 0.1% transaction friction into the continuous MDP reward function in order to empirically resolve the implementation gap [8],[14].Endogenous Safe-Haven Hedging ("CASH"): Embedding a zero-return and zero-volatility liquid asset into the action space, allowing the agent to dynamically de-risk during market crashes [10], [13].Rigorous Empirical Benchmarking: Out-of-sample evaluation of the DRL-PPO against LSTM, RF, and XGBoost on financial ratios, Maximum Drawdown, and CAPM Jensen's Alpha/Beta [8],[10].

The investment space in decentralized finance and cryptocurrencies is defined by an extremely high level of volatility, high-dimensional non-linear relationships, and non-stationary return distributions that are prone to abrupt regime changes [1], [2]. In such an inherently stochastic environment, risk management and capital preservation become crucial for quantitative investments [2], [3]. Traditionally, portfolio construction has been based on the Modern Portfolio Theory (MPT), which was developed by Harry Markowitz back in 1952, whereby the portfolio weights are chosen according to the mean-variance approach to maximize expected returns at a targeted variance [4], [5]. Nevertheless, the Classical MPT and its various econometric extensions are all based on several fundamental assumptions of stationarity of asset prices dynamics, constant correlation matrices, and multivariate normally distributed returns [4], [6]. Indeed, empirical research shows that those parametric assumptions fall apart during the times of digital asset liquidity crisis [2], [7]. Under these circumstances, the optimization process becomes inefficient enough to make the simple diversification approach (1/N) superior to mean-variance methods [7], [8].

To address such issues, the methods of machine learning (ML) and deep learning (DL) have been introduced for modeling such complicated financial patterns [9],[10]. RNN models such as LSTM are commonly applied to model such dependencies, lags, and memory effects within the financial time-series, which are characterized by noise [3], [11]. In addition, ensemble tree models, including Random Forest (RF), LightGBM, and XGBoost have proven to be highly resilient to isolation of complicated non-linear interactions between features [12], [13].

Despite the progress in methodologies, traditional supervised machine learning frameworks suffer from an intrinsic vulnerability in their implementation, often referred to as the implementation gap [14],[15]. Traditional pipelines implement a disjointed "predict-then-optimize" framework, training predictive models purely on statistical point-forecast loss minimization (Mean Squared Error or Mean Absolute Error) [3],[14]. As statistical loss functions are disconnected from reality of market microstructure, small directional forecasting errors trigger overzealous trading signals [14],[15]. With proportion of transaction fees ranging between 0.1% and 0.25% per each portfolio re-allocation step, such over-trading quickly leads to gross alpha erosion. As a result, traditional unsupervised trading strategies are frequently subject to drastic capital depreciation, turning initially double-digit gross alpha into flat or negative net return [16]. Moreover, traditional supervised models exhibit strong market bias and are unable to dynamically de-risk portfolios or hedge capital during extended bear markets and liquidity crisis episodes [2],[17].

On the contrary, Deep Reinforcement Learning (DRL) framework resolves the aforementioned issues by recasting the portfolio optimization problem into a Markov Decision Process (MDP) [14], [18], [19]. By omitting the intermediate step of predicting asset prices through regression models, a DRL agent implements a mapping from observed states to optimal portfolio weights through repeated interaction with the environment ($\pi: S \rightarrow A$) [18], [19]. Among continuous-action policy gradient algorithms, Proximal Policy Optimization (PPO) offers robust empirical stability [20], [21]. Through the use of a clipped surrogate loss function, PPO restricts the extent of policy updates at each iteration, preventing disastrous policy degradation in volatile market environment [20], [21].

Most recently, a substantial body of research has demonstrated the importance of designing reward functions in DRL: by replacing unrestricted profit/loss reward functions with adaptive risk-averse reward functions, agents can deliver superior Sharpe ratios (up to 2.47) while effectively reducing maximum drawdown [2], [7]. Additionally, by explicitly incorporating the cost of transaction into the design of the reward function, the agent is able to account for the friction of execution [4], [14], [18]. Penalizing the L_1-norm of weight adjustments between

time-steps teaches the policy network about transaction frictions, enabling it to execute transactions only if expected marginal gains cover the cost of such a transaction [1], [4], [14]. Further reward function formulations employing geodesic models of transaction costs based on information geometry enable the agent to adjust trading frequency based on liquidity fragmentation [1], [22].

In light of the above discussed empirical and methodological challenges, this study develops a friction-aware portfolio optimization framework implementing PPO on the top ten largest cryptocurrencies by market cap. The main contributions of this paper include:

- Portfolio Optimization with Explicit Transaction Cost Penalty ("Friction-Aware"): Implementation of the 0.1% transaction cost per trade at each rebalance step into the reward function of MDP [4], [14] and bridging the implementation gap empirically.
- Risk Aversion and Dynamic Safe Haven Allocation ("CASH"): Implementation of a risk-free liquid asset with a zero-return rate ("CASH") into the action space, offering the agent an endogenous risk-mitigation strategy by allowing it to dynamically de-risk its portfolio [2], [18].
- Comparative Evaluations of Models: Rigorous comparative out-of-sample benchmarking of PPO policy against leading supervised machine learning architectures (LSTM, XGBoost, Random Forest) in terms of financial metrics, Jensen's Alpha, Beta and Maximum Drawdowns under equal friction constraints [4], [14].

3.2 Materials and methods

Creating an effective framework for the management of dynamic cryptocurrency portfolios needs to have a sequential architecture that can convert data obtained from the market to efficient portfolio allocations (Fig. 1). The current paper uses a four-step methodology framework that involves extraction and preprocessing of data, calculation of technical indicators, training the model using Machine and Deep Learning techniques, and portfolio optimization.

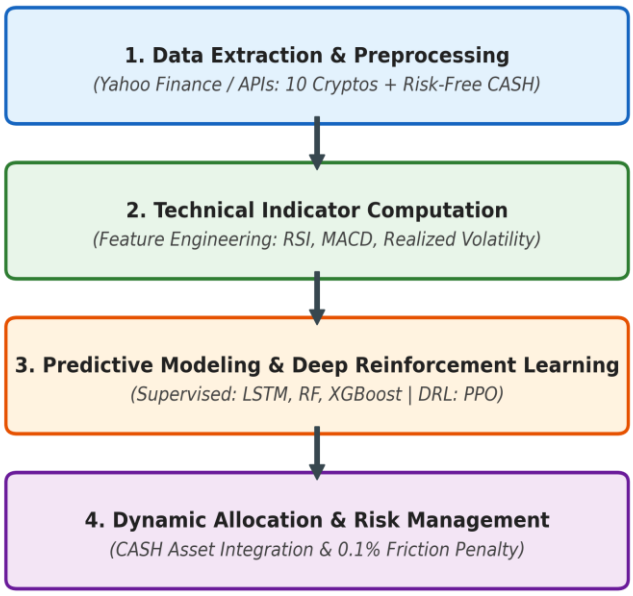


Fig 1. End-to-End Methodology Architecture for Dynamic Cryptocurrency Portfolio Management and Risk Management.

3.1 Data Extraction and Investment Universe

A Data Collection The investment universe is comprised of ten liquid cryptocurrencies chosen by liquidity and market capitalization: Cardano (ADA), Binance Coin (BNB), Bitcoin (BTC), Dogecoin (DOGE), Polkadot (DOT), Ethereum (ETH), Litecoin (LTC), Solana (SOL), TRON (TRX), and Ripple (XRP) [10], [16]. A liquid asset with zero return and zero volatility ("CASH", $E[r_{cash}] = 0$, $\sigma_{cash} = 0$) is used for endogenous hedging during market crashes [16], [17]. The daily close prices series for each of the assets from January 2021 to January 2024 is collected via API with the use of yfinance [10]. Continuous logarithmic daily return for asset i at time t is calculated as formula (1).

$$r_{i,t} = \ln\left(\frac{P_{i,t}}{P_{i,t-1}}\right) \quad (1)$$

3.2 Feature Engineering and Technical Indicators

For capturing the momentum, overbought/oversold status, and volatility, the technical indicators are calculated for every constituent security [18]: Relative Strength Index (RSI): Price Velocity within an n-day period [18]

- Relative Relative Strength Index (RSI): Price Velocity within an n-day period

$$\text{RSI}_t = 100 - \frac{100}{1+\frac{\text{EMA}(\max(P_t - P_{t-1}, 0), n)}{\text{EMA}(\max(P_{t-1} - P_t, 0), n)}} \quad (2)$$

Where EMA stands for Exponential Moving Average.

- Moving Average Convergence Divergence (MACD): Used to reveal the relation between two exponential moving averages and determine trend changes [18].

$$\text{MACD}_t = \text{EMA}(P, 12) - \text{EMA}(P, 26) \text{ with } \text{Signal}_t = \text{EMA}(\text{MACD}, 9) \quad (3)$$

- Historical Annualized Volatility: Accounts for time-dependent volatility clustering over a rolling window W = 20 trading days, annualized according to the 365-day convention of cryptocurrencies [19], [20].

$$\sigma_{i,t} = \sqrt{\frac{1}{W}\sum_{k=0}^{W-1}\left(r_{i,t-k} - \bar{r}_{i,t}\right)^2} \times \sqrt{365} \quad (4)$$

Historical Annualized Volatility $\sigma_{i,t}$.

3.3. Predictive modeling and Deep Reinforcement Learning (DRL-PPO)

In the experiment, there is a comparison between two different methodological approaches [5], [6], [7]. Classic supervised regression models (Predict, then Optimize): Subsequent to point forecasting, the weights of portfolios are determined through the use of Sequential Least Squares Programming (SLSQP) that maximizes Sharpe ratio under simplex constraints of long-only $\sum_{i=1}^{11} w_{i,t} = 1$ and $0 \leq w_{i,t} \leq 1$ for all i.

2.3.1. Long Short-Term Memory (LSTM)

LSTM Neural Network**:** Captures long-term sequential dependencies and temporal autocorrelation through input, forget, and output gating mechanisms [5], [21]: The recurrent neural network architecture is built for modeling the long-term temporal relationships in the time series of the financial market. The algorithm captures the momentum of the market by analyzing the past data windows.

$$f_t = \sigma\left(W_f \cdot [h_{t-1}, x_t] + b_f\right) \quad (5)$$

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \quad (6)$$

$$\tilde{C}_t = \tanh(W_c \cdot [h_{t-1}, x_t] + b_c) \quad (7)$$

$$C_t = f_t \odot C_{t-1} + i_t \odot \tilde{C}_t \quad (8)$$

$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) \quad (9)$$

$$h_t = o_t \odot \tanh(C_t) \quad (10)$$

Forget input and output gates, respectively are denoted by f_t , i_t and o_t, C_t denotes the cell state, and h_t denotes the hidden vector.

2.3.2. Random Forest (RF) and XGBoost

This type of tree-based ensemble models is used in order to filter out structural noise from the market and predict future returns. Thanks to using methods such as bootstrapping, the Random Forest algorithm uncovers nonlinear relations that may not be captured by neural networks. [6], [7],[22],[23].

Quadratic loss minimization decision trees, where the goal is to predict future $\hat{r}_{i,t+1}$ using technical indicators.

$$\mathcal{L}_{\text{MSE}} = \sum_t \left(r_{i,t+1} - \hat{r}_{i,t+1}\right)^2 \quad (11)$$

2.3.3. Deep Reinforcement Learning Algorithm DRL-PPO

The DRL framework does not utilize classical models that solely depend on predicting directional movement of prices ("predict-then-optimize"). Instead, the DRL framework treats portfolio management as a Markov Decision Process (MDP) [10],[24]. and within this framework, the autonomous agent learns to allocate capital based on its continuous interaction with the environment (trial-and-error).

- **Proximal Policy Optimization (PPO)**

At the heart of this framework lies PPO, which is a state-of-the-art actor-critic algorithm designed by OpenAI. PPO works exceptionally well in stochastic and noisy environments such as financial markets. This algorithm uses a clipped surrogate objective function that restricts the magnitude of policy update at each iteration. Thus, PPO guarantees monotonic improvement while preventing destructive updates to the policy, giving this model the stability needed to operate during extreme volatility of cryptocurrency prices.

- **State Space S_t**

At every point in time t, the agent gets information about the environment in the form of the state space S_t [10], [25]. To make sure that the agent has enough information, S_t contains three different data dimensions:

Price Dynamics: The normalized historical logarithmic returns of the targeted assets. Technical Context.

Technical indicators (such as MACD, RSI, and GARCH volatility) extracted from the market that give the agent information on trends, overbought/oversold state, and volatility clustering.

Temporal Portfolio State: Allocation weights of the portfolio in the previous period S_{t-1}Remembering the portfolio state is crucial because the agent needs this information to calculate the portfolio turnover and estimate the cost of transactions before taking any actions.

- **Action Space A_t and CASH**

Based on the state of the environment S_t, the policy (or actor) network produces an action A_t[25],[26]. In the continuous action space, A_t is the target portfolio weights. A softmax activation function in the last layer ensures that all weights will be positive (because the model cannot sell short) and sum to 1.

Notably, apart from the targeted cryptocurrencies, the action space includes the risk-free asset called CASH (for example, fiat-pegged stablecoin with zero interest rate). Including the CASH gives the agent the endogenous means to hedge itself. Instead of being forced to stay fully invested when the market collapses, the agent learns to detect changes in the regime. In bear markets or liquidity crises, the PPO agent can adopt a "flight-to-safety" strategy and move most of the weights into the CASH position.

- **Reward Function**

Without Market Frictions The gross return on the portfolio R_t^{raw} obtained from the allocation vector $\mathbf{w}_t = [w_{1,t}, w_{2,t}, \dots, w_{11,t}]^T$ is formulated at time period t in the following manner

$R_t^{\text{raw}} = \sum_{i=1}^{11} w_{i,t} \cdot r_{i,t} = \mathbf{w}_t^T \mathbf{r}_t$ (12)

$\sum_{i=1}^{11} w_{i,t} = 1$ and $0 \le w_{i,t} \le 1$ for all i.

Where N+1 includes the number of cryptocurrencies and cash

In order to avoid excessive trading and represent market conditions, the reward function R_t is formulated with an additional transaction cost λ

$R_t = \ln(1 + R_t^{\text{raw}}) - \lambda \sum_{i=1}^{11} |w_{i,t} - w_{i,t-1}|$ (13)

The objective of portfolio management is to try to maximize the return and minimize the risk. The Sharpe ratio is measured using formula (14), whereby the numerator is the return and the denominator is the risk. Thus, maximizing the Sharpe ratio means maximizing the return and minimizing the risk. In order to consider the transaction cost, we will subtract the cost in each of our calculations leading to the reward function measured using equation (15).

$$S_t = \frac{R_{P,t} - R_{rf}}{\sigma_P} \quad (14)$$

$$R_t = \frac{R_{P,t} - R_{rf}}{\sigma_P} - \lambda \sum_{i=1}^{N} |w_{i,t} - w_{i,t-1}| \quad (15)$$

Where:

- $w_{i,t}$ is the weight of the i-th security at time t,
- $w_{i,t-1}$ is its weight at time $t-1$,
- λ is the cost of transaction for each rebalance.
- $R_{P,t}$ portfolio return at time t
- R_{rf} free risk
- σ_P portfolio risk

2.4 Measuring portfolio performance

Changes in wealth out of sample are measured through the compound return on net investment: $W_t = W_0 \prod_{\tau=1}^{t}(1 + R_{p,\tau})$ [28].

- Sharpe & Sortino Ratios: Calculated for excess returns over total risk σ_{ann} and downside risk (σ_{down}), respectively. Risk σ_{down}.

  $$\text{SR} = \frac{R_{\text{ann}} - R_f}{\sigma_{\text{ann}}} \quad , \quad \text{Sortino} = \frac{R_{\text{ann}} - R_f}{\sigma_{\text{down}}} \qquad (16)$$

- Maximum Drawdown (MDD) and Calmar Ratio: Measuring tail risk and efficiency of capital recovery [28]:

  $$\text{MDD} = \min_{t \in [1,T]} \left(\frac{W_t - \max_{\tau \le t} W_\tau}{\max_{\tau \le t} W_\tau} \right), \quad \text{Calmar} = \frac{R_{\text{ann}}}{|\text{MDD}|} \qquad (17)$$

- CAPM –alpha and beta attribution: Separating performance into systematic risk $\beta$and added value from strategy α relative to the benchmark $R_{m,t}$ [30]:

  $$R_{p,t} - R_{f,t} = \alpha + \beta\left(R_{m,t} - R_{f,t}\right) + \epsilon_t \qquad (18)$$

3.3 Result and discussion

3.1 Statistical Characterization of the Investment Universe

The cross-asset dynamics of the cryptocurrency portfolio from 2021 to 2024 display very high volatility clustering and correlations [31]. The cumulative return paths of each asset in Fig. 2 illustrate the high volatility dynamics of the 2021 bull period, followed by the 2022 bear period and the recovery in late 2023 [31]. Cross-asset Pearson correlation coefficient matrix for the given assets (Fig.3) displays very high positive dependencies between digital currencies with the correlation coefficients being often higher than 0.60 (for instance, BTC-ETH: 0.81, ETH-LTC: 0.79) [31], [32]. Such extensive interdependencies structurally hinder passive diversification gains [2], [32]. Conversely, the CASH asset has no correlation with other assets ($\boldsymbol{\rho} = \mathbf{0}$).

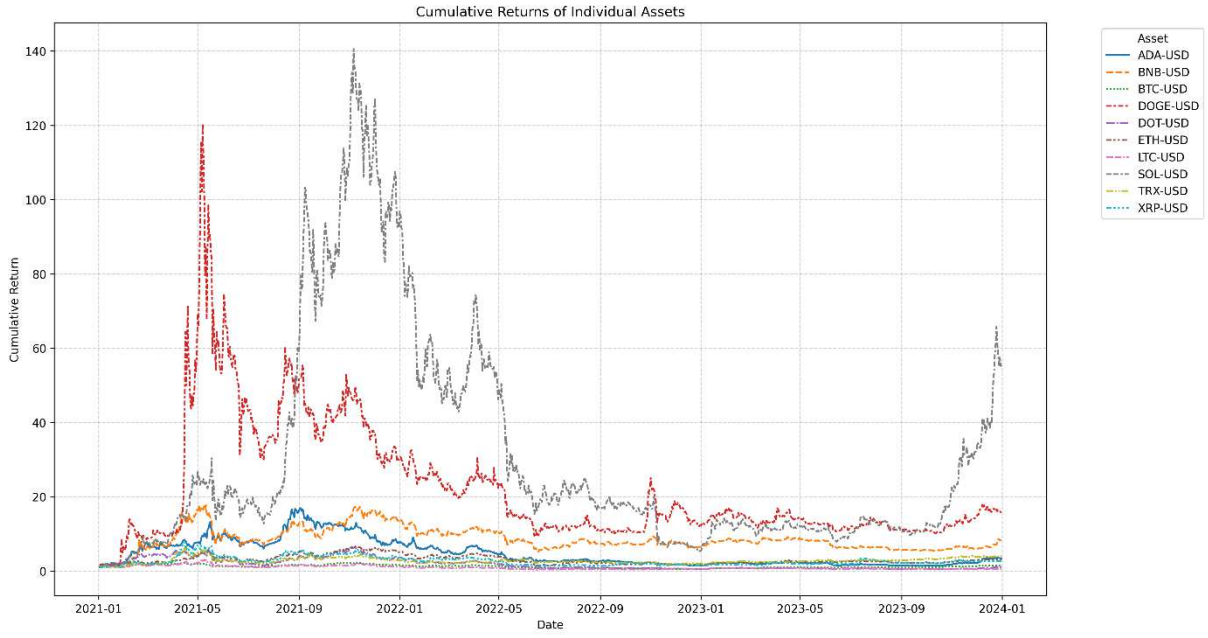


Fig. 2 return for Individual asset (2021-2024)

Fig 3 shows the correlation matrix for returns on the asset. It is evident that there is an extremely strong positive and significant correlation between the returns of the assets of financial instruments in cryptocurrency markets in all cryptocurrencies, thus structurally constraining the efficiency of passive diversification. The only safe-haven asset that is immune to market risk is CASH, which does not have any correlation with risky assets.

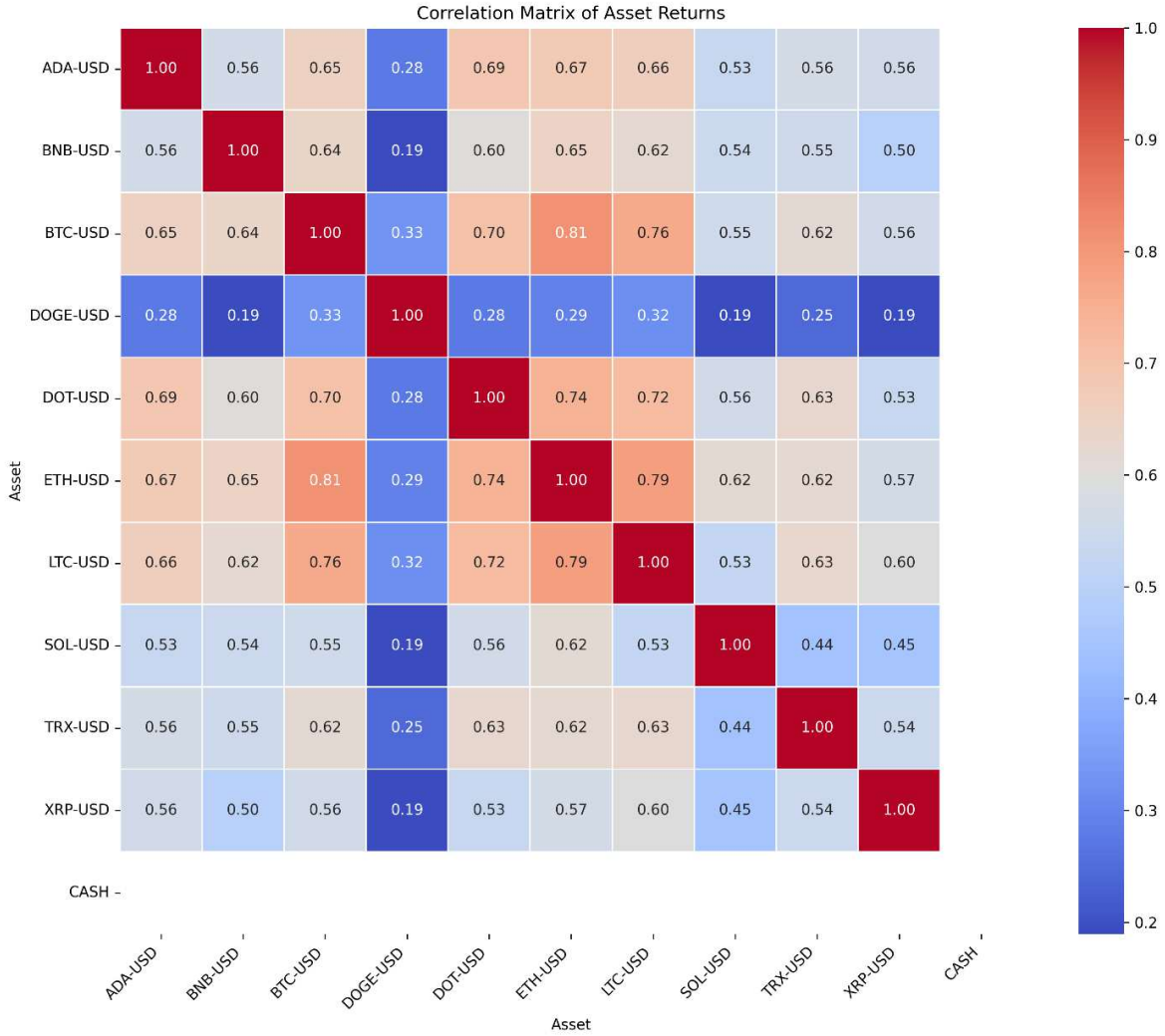


Fig. 3 Correlation matrix of returns on assets

3.2 Comparative Analysis of Global Performances

The performance indicators measured for the considered models on an out-of-sample basis under rebalancing friction of 0.1% are presented in Table 1.

Table 1 Performance Comparison Criteria (Test Duration)

Model	Annual return (%)	Annual risk (%)	Sharpe ratio	Sortino Ratio	Calamar Ratio	Max Drawdown (%)
DRL (PPO)	**37.22**	**42.58**	**0.87**	**1.34**	**1.30**	**-28.74**
LSTM	6.75	51.94	0.13	0.18	0.20	-33.70
Random Forest	23.54	55.96	0.42	0.62	0.80	-29.29
XGBoost	-1.48	52.08	-0.03	-0.04	-0.05	-30.75

Results of the empirical study showcase the obvious superiority of the DRL-PPO method over other supervised approaches: Efficiency: DRL-PPO is the most efficient agent according to the risk-adjusted metrics with an average annual profit of 37.22%, Sharpe Ratio of 0.87, Sortino Ratio of 1.34 and Calmar Ratio of 1.30 [10]. The high Sortino Ratio indicates that the policy of PPO successfully reduces the downside deviation while preserving the positive return [13],[33].Capital Protection : The Maximum Drawdown of the PPO method amounts to -28.74%, outperforming other algorithms such as Random Forest (-29.29%), XGBoost (-30.75%), LSTM (-33.70%) [10],[34].Neutrality to Systemic Risk and Positive $\boldsymbol{\alpha}$ Generation: PPO achieves the near-to-zero Beta $\boldsymbol{\beta}$= -0.08), becoming the agent of active market neutrality and generating the positive Jensen's Alpha of +46.41% [10],[35]. On the contrary, other supervised approaches demonstrate Betas near unity ($\boldsymbol{\beta}$ =0.96-1.04) and negative $\boldsymbol{\alpha}$ even down to -119.12.

3.3 Out-of-Sample Wealth Pathways

In Fig. 4, the wealth plots adjusted for transaction costs exhibit a dual-phase market behavior [10], [36]:

Bearish Compression Period (June to October 2023): The DRL algorithm (blue plot) is capable of maintaining the portfolio's wealth value close to 0.80 without experiencing large drawdowns through active redistribution of funds towards the CASH stock [10], [16]. However, supervised learners are subject to strong depreciation: Random Forest reaches 0.73 in June, and LSTM reaches the lowest value on the panel below 0.69 in October [10].

Bullish Expansion Period (November 2023 to January 2024): The DRL algorithm maintains an increasing trend, exceeding 1.0 in November, reaching more than 1.24, and finishing first around 1.20 [10]. Random Forest increases up to ~1.14 with high volatility, while LSTM stays far behind with ~1.04 [10]. Meanwhile, XGBoost (red plot) fails to perform well due to constant rebalancing and ends in the red zone around 0.99 [8], [10].

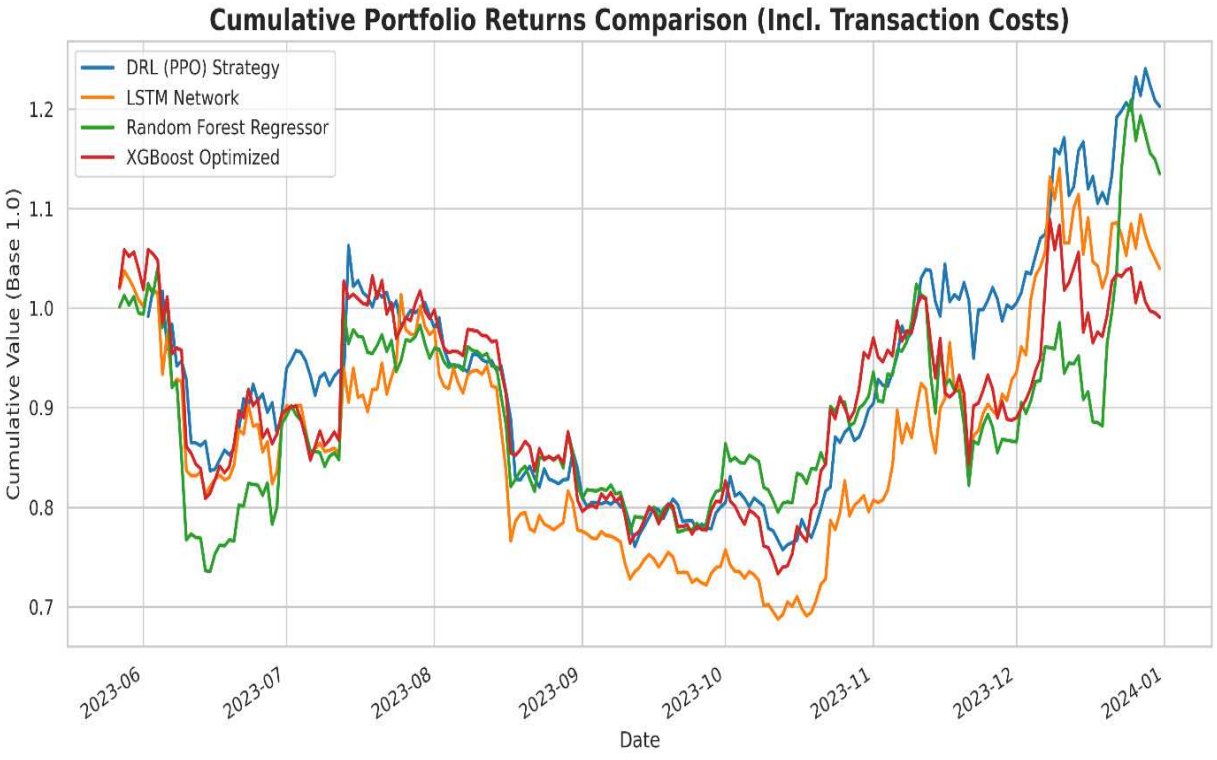


Fig.4 Cumulative Out-of-Sample Portfolio Returns with 0.1% Transactions Costs.

3.4. Mechanistic Understanding and Execution Gaps

The boxplot for daily returns (Fig 5), Beta coefficients (Fig 6), and Jensen's Alpha distributions (Fig 7) reveal the mechanisms behind this performance gap [10], [36]: Decision-aware strategy versus predict-then-optimize: In supervised learning-based approaches, the optimization objective is based on an unsupervised loss function without any link to finance [8], [37]. Contrary to this, DRL directly optimizes risk-adjusted returns in the environment, without forecasting errors [10], [37]. Turnover Costs: Rebalancing each day under statistical methods incurs compounded transaction costs [8], [14]. PPO's reward function includes the penalty on the L_1-norm of the weight changes, and acts only if the expected gain compensates for them [14], [36].

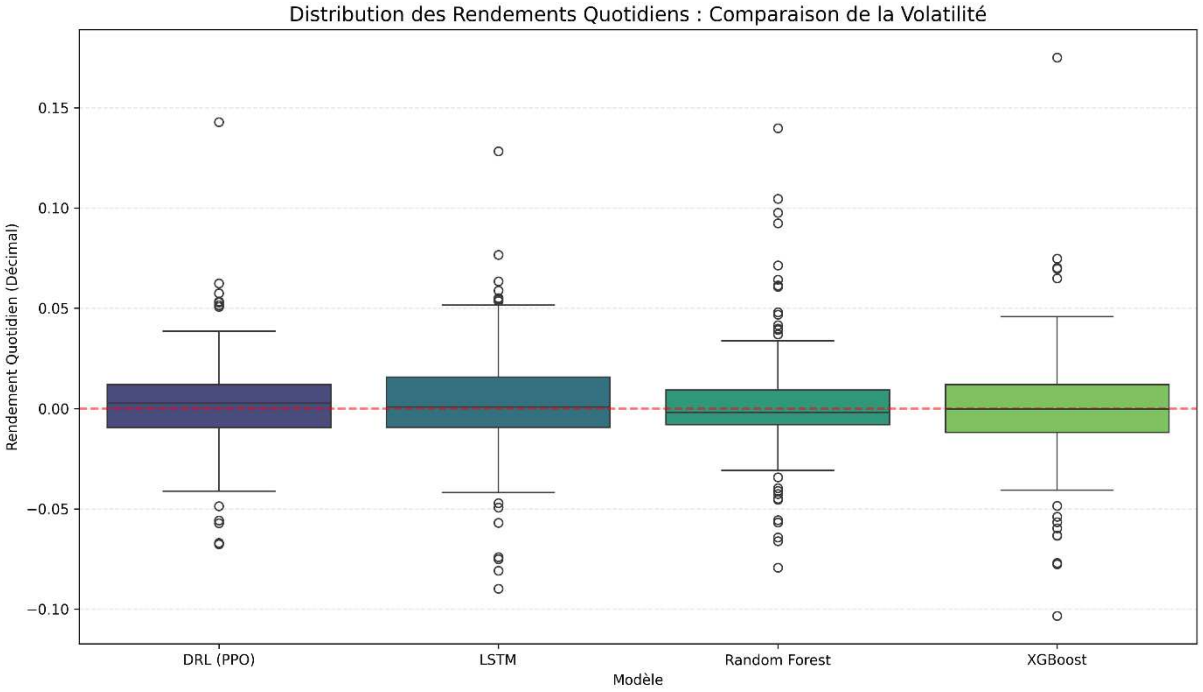


Fig 5. Distribution of Daily Returns and Outliers Comparison for Portfolio Strategies.

3.5 Market Risk and Attribution of Alpha through Different Portfolio Strategies

Performance Attribution: The market Beta and Jensen's Alpha analysis shows two completely different types of behavioral strategies used in the investigated models:

- The Beta coefficient of the Random Forest and XGBoost models is larger than or equal to one (1.04 for both models; see Fig 7). This implies that these models use a purely passive strategy that follows the market price and is exposed to the systematic risk of the cryptocurrency market. Additionally, the large negative values of the Jensen's Alpha (about -100% to -119%; see Fig 6) prove that they constantly destroy economic value as compared to an equally weighted market index.
- The PPO DRL model has almost zero Beta with -0.08 (Fig 5). This means that it uses a decorrelation strategy typical for market neutral approaches. More importantly, this is the only tested framework that has a significantly positive Jensen's Alpha of +46.41% (Fig 6). This is evidence that the better performance of the PPO model cannot be due to passive investment in bullish markets but due to actual alpha generation.

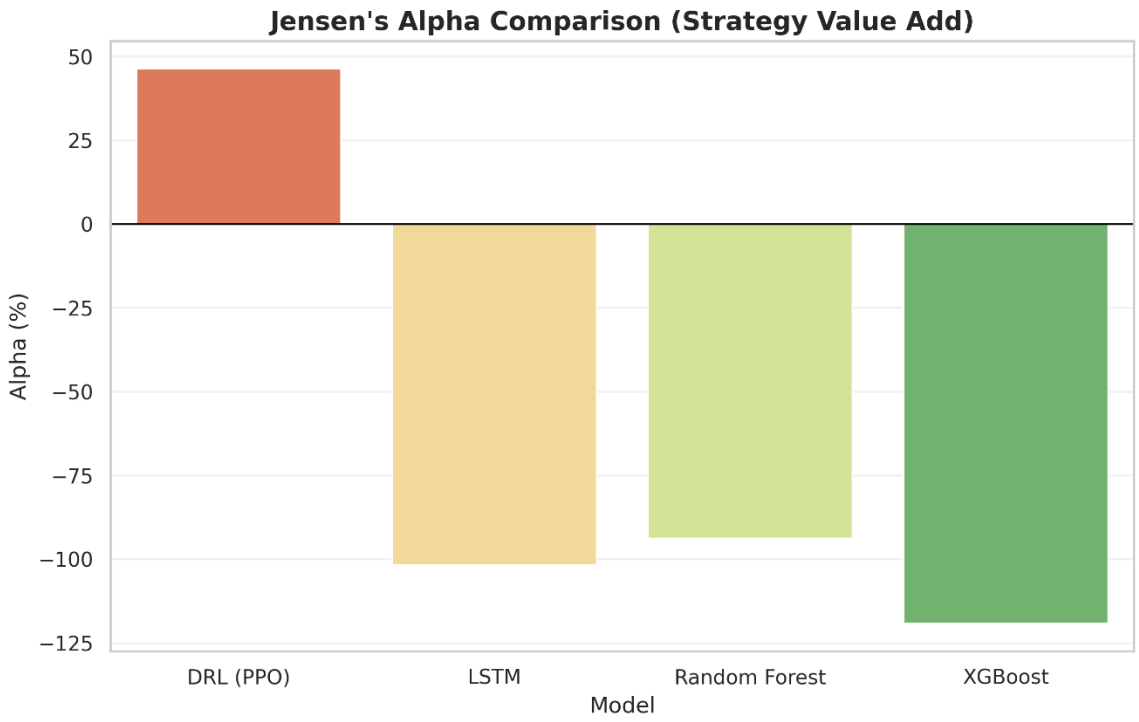


Fig 6: Alpha of Jensen Compared (Strategy Value Addition).

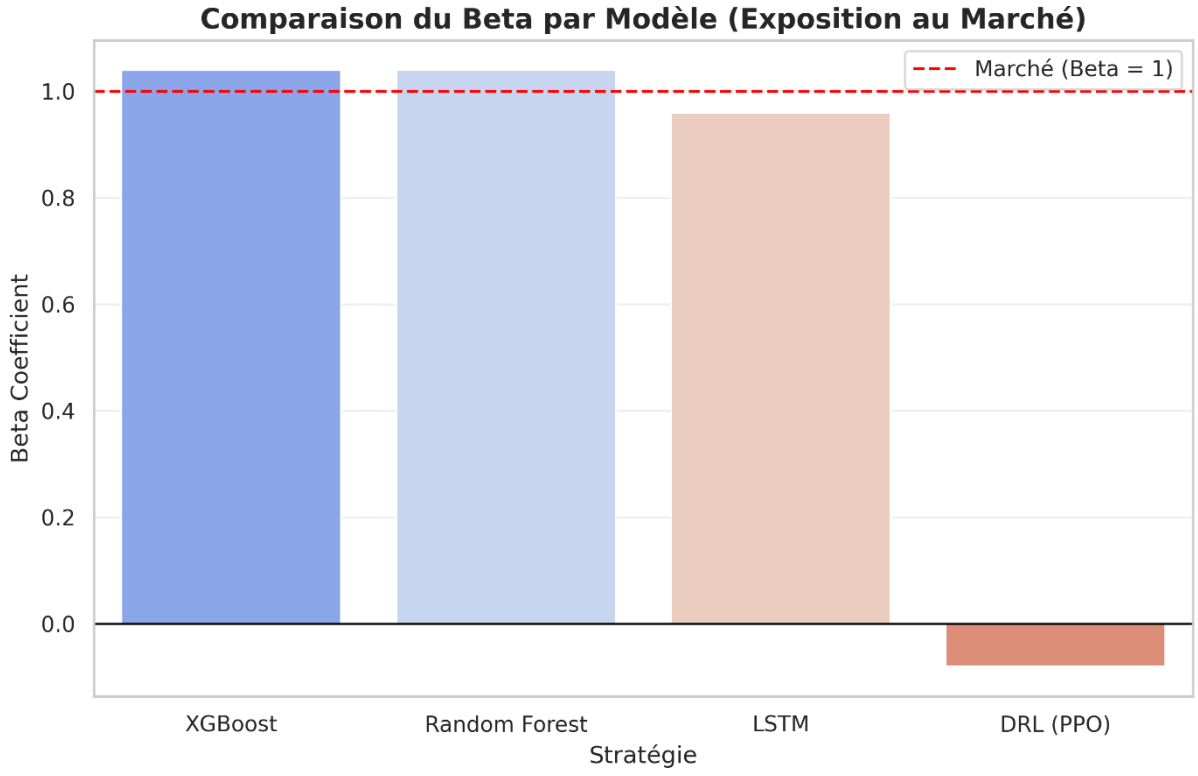


Fig 7. Comparison of Market Beta in Various Portfolio Allocations.

3.6 Limitations and future research

Despite the high empirical results achieved by the DRL-PPO framework, some limitations should be considered [10]:

Purely Exogenous State Space: The state space is constructed using only price movements and technical indicators; any exogenous components (regulatory actions, macroeconomic events, social media analysis) are ignored [38], [39].

Execution Friction as a Static Parameter: Execution costs as 0.1% are considered static, whereas the actual costs depend on market depth and order flow dynamics [14], [36].

Trading Based on Daily Data Points: Trading at the end of each day ignores the intraday liquidity dynamics [16], [37].

For future research, we will investigate:

Multimodal Data Aggregation: Utilization of multimodal data from financial text analysis using the Fin BERT model [38], [39], on-chain metrics (such as exchange flows and hash rate), in addition to the price information in the state [38].

Tail Risk Asymmetric Reward Function: Including conditional Value-at-Risk (CVaR)/Expected Shortfall in the optimization process of policy gradients to better handle fat tails in returns [40].

Multi-Agent Systems & Online Learning: Using hierarchical MARL algorithms for collaborative allocation and hedging, as well as the use of online walk-forward learning to adapt to changing market conditions [41].

5 Conclusion

This work presents experimental results of the effectiveness of Deep Reinforcement Learning (DRL-PPO) compared to traditional supervised regression models for the management of cryptocurrency portfolios dynamically [10], [42]. Thanks to the built-in ability to consider the risk management, transaction costs, and safe-haven investment, the Sharpe Ratio of PPO is 0.87, with the overall return of 37.22%, without compromising the market neutrality (β = -0.08$) and having the positive value of Jensen's Alpha (+46.41%) [10]. In turn, traditional methods have an implementation problem because of the point-prediction errors and the cost of turnover decreasing the total capital [8], [42].

.

Declarations

Author contributions

Conceptualization and methodology, K.E.K and D.E.; software, formal analysis, data curation, and visualization, D.E.; validation and interpretation, D.E., K.E.K., N.F., M.E.M., S.M., B.B., K.E.F., M.E.F., and M.T.B.; writing - original draft preparation, D.E.; writing - review and editing, all authors; supervision, K.E.K. and M.T.B. All authors have read and agreed to the final version of the manuscript.

Funding

This research received no external funding.

Data availability

The data sets utilized in this study were collected from the publicly accessible financial databases. Historical data on price and volume of the 10 selected cryptocurrency assets

such as Cardano (ADA-USD), Binance Coin (BNB-USD), Bitcoin (BTC-USD), Dogecoin (DOGE-USD), Polkadot (DOT-USD), Ethereum (ETH-USD), Litecoin (LTC-USD), Solana (SOL-USD), TRON (TRX-USD), and Ripple (XRP-USD) for the period from January 1, 2021, to January 1, 2024, were collected through the Yahoo Finance API via the yfinance open source package. It is supposed that the risk-free liquid asset (CASH) has a zero daily return and a zero variance. In order to ensure complete transparency and replicability, all collected data sets, returns calculations, and experimental Python scripts are publicly available at: https://github.com/kamal-107/deep-reinforcement-learning-Cryptos-portfolio-optimization-de.git

Ethics statement

Ethics approval and consent to participate: Not applicable. The study uses publicly available financial-market data and does not involve human participants, animals, or identifiable personal data.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

1. H. Markowitz, "Portfolio selection," *The Journal of Finance*, vol. 7, no. 1, pp. 77–91, Mar. 1952, doi: 10.1111/j.1540-6261. 1952.tb01525. x.
2. V. DeMiguel, L. Garlappi, and R. Uppal, "Optimal versus naive diversification: How inefficient is the 1/N portfolio strategy?" *The Review of Financial Studies*, vol. 22, no. 5, pp. 1915–1953, May 2009, doi: 10.1093/rfs/hhm075.
4. S. Corbet, B. Lucey, A. Urquhart, and L. Yarovaya, "Cryptocurrencies as a financial asset: A systematic analysis," *International Review of Financial Analysis*, vol. 62, pp. 182–199, Mar. 2019, doi: 10.1016/j.irfa.2018.09.003.
5. F. Fang, C. Ventre, M. Basios, L. Kanthan, D. Martinez-Rego, F. Wu, and L. Li, "Cryptocurrency trading: A comprehensive survey," *Financial Innovation*, vol. 8, no. 1, art. no. 13, Jan. 2022, doi: 10.1186/s40854-021-00321-6.
6. T. Fischer and C. Krauss, "Deep learning with long short-term memory networks for financial market predictions," *European Journal of Operational Research*, vol. 270, no. 2, pp. 654–669, Oct. 2018, doi: 10.1016/j.ejor.2017.11.054.
7. S. Gu, B. Kelly, and D. Xiu, "Empirical asset pricing via machine learning," *The Review of Financial Studies*, vol. 33, no. 5, pp. 2223–2273, May 2020, doi: 10.1093/rfs/hhaa009.
8. T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discov. Data Min.*, Aug. 2016, pp. 785–794, doi: 10.1145/2939672.2939785.
9. A. Bysik and R. Ślepaczuk, "Machine learning-based Bitcoin trading under transaction costs: Evidence from walk-forward forecasting," *arXiv preprint arXiv:2606.00060*, 2026.

10. A. Millea, "Deep reinforcement learning for trading - A critical survey," *Data*, vol. 6, no. 11, art. no. 119, Nov. 2021, doi: 10.3390/data6110119.

11. J. Moody and M. Saffell, "Learning to trade via direct reinforcement," *IEEE Transactions on Neural Networks*, vol. 12, no. 4, pp. 875–889, Jul. 2001, doi: 10.1109/72.935097.

12. J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal policy optimization algorithms," *arXiv preprint arXiv:1707.06347*, 2017.

13. X. Wang, Y. Chen, W. Lu, X. Wang, and Z. Qin, "Deep reinforcement learning: A survey," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 35, no. 4, pp. 5064–5078, Apr. 2024, doi: 10.1109/TNNLS.2022.3207346.

14. E. H. Khujamatov, K. Ismanov, O. U. Mallaev, and O. Sattarov, "Optimizing crypto-trading performance: A comparative analysis of innovative reward functions in reinforcement learning models," *Mathematics*, vol. 14, no. 5, art. no. 794, Feb. 2026, doi: 10.3390/math14050794.

15. Y. Zhang, P. Zhao, Q. Wu, B. Li, J. Huang, and M. Tan, "Cost-sensitive portfolio selection via deep reinforcement learning," *IEEE Transactions on Knowledge and Data Engineering*, vol. 34, no. 10, pp. 4710–4722, Oct. 2022, doi: 10.1109/TKDE.2020.3044431.

16. N. D. Moroke, "Deep reinforcement learning for cryptocurrency portfolio management: A free-energy PPO framework with geodesic transaction costs," *Preprints.org*, 2026, doi: 10.20944/preprints 202601.0112.v1.

17. S. R. Sahu, J. H. Ochoa Vázquez, A. F. Ramírez, and J. M. Kim, "Analyzing portfolio optimization in cryptocurrency markets: A comparative study of short-term investment strategies using hourly data approach," *Research in International Business and Finance*, vol. 72, art. no. 102514, Dec. 2024, doi: 10.1016/j.ribaf.2024.102514.

18. T. Conlon and R. McGee, "Safe haven or risky hazard? Bitcoin during the COVID-19 bear market," *Finance Research Letters*, vol. 35, art. no. 101607, Jul. 2020, doi: 10.1016/j.frl.2020.101607.

19. S. Mostafavi and S. Hooman, "Key technical indicators for stock market prediction," *Machine Learning with Applications*, vol. 22, art. no. 100478, Dec. 2025, doi: 10.1016/j.mlwa.2025.100478.

20. R. Chen, "ARIMA vs. machine learning in portfolio return forecasting: A comparative study integrating GARCH-based volatility estimation," in *Proc. 13th Int. Conf. Data Sci. Technol. Appl. (DATA)*, 2024, pp. 120–128, doi: 10.5220/0012756000003756.

21. T. Bollerslev, "Generalized autoregressive conditional heteroskedasticity," *Journal of Econometrics*, vol. 31, no. 3, pp. 307–327, Apr. 1986, doi: 10.1016/0304-4076(86)90063-1.

22. W. Bao, J. Yue, and Y. Rao, "A deep learning framework for financial time series using stacked autoencoders and long-short term memory," *PLOS ONE*, vol. 12, no. 7, art. no. e0180944, Jul. 2017, doi: 10.1371/journal.pone.0180944.

23. M. Stawarz and M. D. Stasiak, "Determining multi-class trading signals for Bitcoin: A comparative study of XGBoost, LightGBM, and Random Forest," in *Proc. 33rd Int. Conf. Inf. Syst. Dev. (ISD 2025)*, 2025, pp. 1–12.

24. H. Kim, "Mean-variance portfolio optimization with stock return prediction using XGBoost," *Economic Computation and Economic Cybernetics Studies and Research*, vol. 55, no. 4, pp. 5–20, Dec. 2021, doi: 10.24818/18423264/55.4.21.01.

25. V. Mnih *et al.*, "Human-level control through deep reinforcement learning," *Nature*, vol. 518, no. 7540, pp. 529–533, Feb. 2015, doi: 10.1038/nature14236.

26. R. Yan, J. Jin, and K. Han, "Reinforcement learning for deep portfolio optimization," *Electronic Research Archive*, vol. 32, no. 9, pp. 5176–5200, Sep. 2024, doi: 10.3934/era.2024237.

27. F. Soleymani and E. Paquet, "Financial portfolio management using deep reinforcement learning," *Expert Systems with Applications*, vol. 182, art. no. 115278, Nov. 2021, doi: 10.1016/j.eswa.2021.115278.

28. K. Paykan, "Cryptocurrency portfolio management with reinforcement learning: Soft Actor-Critic and Deep Deterministic Policy Gradient algorithms," *arXiv preprint arXiv:2511.20678*, 2025.

29. F. A. Sortino and L. N. Price, "Performance measurement in a downside risk framework," *The Journal of Investing*, vol. 3, no. 3, pp. 59–64, Aug. 1994, doi: 10.3905/joi.3.3.59.

30. W. F. Sharpe, "Mutual fund performance," *The Journal of Business*, vol. 39, no. 1, pp. 119–138, Jan. 1966, doi: 10.1086/294846.

31. E. F. Fama and K. R. French, "Common risk factors in the returns on stocks and bonds," *Journal of Financial Economics*, vol. 33, no. 1, pp. 3–56, Feb. 1993, doi: 10.1016/0304-405X (93)90023-5.

32. A. Bouteska, M. Z. Abedin, P. Hajek, and K. Yuan, "Cryptocurrency price forecasting – A comparative analysis of ensemble learning and deep learning methods," *International Review of Financial Analysis*, vol. 92, art. no. 103055, Mar. 2024, doi: 10.1016/j.irfa.2023.103055.

33. L. Del Nero and P. Giudici, "Machine learning models to predict stock market spillovers," *Finance Research Letters*, vol. 86, art. no. 108508, Dec. 2025, doi: 10.1016/j.frl.2025.108508.

34. T. Bauman, L. Mrčela, S. Goluža, and Z. Kostanjčar, "A deep learning approach to goal-based portfolio optimization in non-stationary environments," *IEEE Access*, vol. 13, art. no. 3588247, pp. 31200–31215, 2025, doi: 10.1109/ACCESS.2025.3588247.

35. M. Azamjon, O. Sattarov, and D. Na, "Enhanced Bitcoin price direction forecasting with DQN," *IEEE Access*, vol. 12, pp. 29093–29112, Feb. 2024, doi: 10.1109/ACCESS.2024.3368291.

36. J. Kevin and P. Yugopuspito, "Hybrid LSTM and PPO networks for dynamic portfolio optimization," *arXiv preprint arXiv:2502.04512*, 2025.

37. C. Öztürk, "Multi-agent deep reinforcement learning with dynamic portfolio weighting: A novel approach to algorithmic trading," *Sakarya University Journal of Computer and Information Sciences*, vol. 9, sp. iss., pp. 591–608, 2026, doi: 10.35377/saucis.09.02.102.

38. S. Makridakis, E. Spiliotis, R. Hollyman, F. Petropoulos, N. Swanson, and A. Gaba, "The M6 forecasting competition: Bridging the gap between forecasting and investment

decisions," *International Journal of Forecasting*, vol. 41, no. 4, pp. 1315–1354, Oct. 2025, doi: 10.1016/j.ijforecast.2024.11.002.

39. A. H. Huang, H. Wang, and Y. Yang, "FinBERT: A large language model for extracting information from financial text," *Contemporary Accounting Research*, vol. 40, no. 2, pp. 806–841, Jun. 2023, doi: 10.1111/1911-3846.12832.
40. F. Wang and M. Xue, "Finance-specific large language models: Advancing sentiment analysis and return prediction with LLaMA 2," *Pacific-Basin Finance Journal*, vol. 90, art. no. 101930, Apr. 2025, doi: 10.1016/j.pacfin.2024.101930.
41. A. Charkhestani and A. Esfahanipour, "Behaviorally informed deep reinforcement learning for portfolio optimization with loss aversion and overconfidence," *Annals of Operations Research*, vol. 338, no. 2, pp. 645–678, Feb. 2026, doi: 10.1007/s10479-025-06489-x.
42. N. Pippas, E. A. Ludvig, and C. Turkay, "The evolution of reinforcement learning in quantitative finance: A survey," *ACM Computing Surveys*, vol. 57, no. 6, art. no. 142, pp. 1–38, May 2025, doi: 10.1145/3712589.
43. H. Gao, H. Ding, and Y. Tong, "Machine learning in business and finance: A literature review and research opportunities," *Financial Innovation*, vol. 10, no. 1, art. no. 86, Sep. 2024, doi: 10.1186/s40854-024-00629-z