Asian Journal of Economics and Banking

Deep reinforcement learning for portfolio allocation: a comparative study on five Vietnamese equities

Worrawat Saijai and Kansuda Pankwaen
Faculty of Economics, Center of Environmental, Social, Governance for Mekong Economies, Chiang Mai University, Chiang Mai, Thailand

Received 25 October 2025
Revised 7 April 2026
25 April 2026
2 May 2026
5 May 2026
30 May 2026
Accepted 5 June 2026

Abstract

Purpose – This study applies deep reinforcement learning (DRL) to multi-asset portfolio optimization in the Vietnamese stock market, aiming to evaluate the performance and stability of different DRL algorithms under emerging market conditions.

Design/methodology/approach – Seven algorithms – A2C, proximal policy optimization (PPO), deep deterministic policy gradient, twin delayed deep deterministic policy gradient (TD3), soft actor-critic (SAC), truncated quantile critics (TQC) and RecurrentPPO – are trained on daily data from January 2018 to December 2024 and evaluated out-of-sample from January 2023 through September 2025 using five liquid equities (SBT.VN, BID.VN, CTG.VN, HPG.VN and VCB.VN). The state representation includes technical indicators (RSI, MACD, SMA, EMA, Bollinger Bands and OBV), as well as risk features such as rolling covariance and a turbulence index.

Findings – PPO achieves the highest annual return (0.1666) and demonstrates the most stable performance. TD3 delivers comparable cumulative growth with higher variability. RecurrentPPO attains the highest Sharpe ratio (1.0461), highlighting the importance of temporal modeling. SAC and TQC produce more conservative but stable outcomes.

Originality/value – This study does not propose a new DRL algorithm; instead, it provides a controlled and reproducible benchmarking framework for comparing multiple DRL models under identical conditions in an emerging market setting.

Keywords Vietnam stock market, Deep reinforcement learning, Portfolio optimization, PPO, RecurrentPPO, Emerging markets

Paper type Research article

1. Introduction

Deep reinforcement learning (RL) has emerged as a powerful framework for portfolio allocation, enabling agents to learn adaptive investment policies directly from price data instead of relying on predefined heuristics or static optimization rules. By combining RL with deep neural networks, DRL agents approximate non-linear, high-dimensional mappings between market states and optimal portfolio weights, supporting robust decision-making under uncertainty (Jiang *et al*., 2017; Deng *et al*., 2016).

Several DRL algorithms have been widely applied in portfolio management, including advantage actor–critic (A2C), proximal policy optimization (PPO) (Schulman *et al*., 2017a), deep deterministic policy gradient (DDPG), twin delayed DDPG (TD3) (Fujimoto *et al*., 2018) and SAC (Haarnoja *et al*., 2018). These models differ by learning strategy (on-vs. off-policy), exploration mechanism and stability trade-off yet share the goal of maximizing long-term return while adapting to stochastic market dynamics. Recent extensions such as truncated

JEL Classification — G11, G17, C45, C63

Funding: This work was supported by the Faculty of Economics, Chiang Mai University.

Asian Journal of Economics and Banking
Emerald Publishing Limited
e-ISSN: 2633-7991
p-ISSN: 2615-9821
DOI 10.1108/AJEB-10-2025-0168

quantile critics (TQC) and RecurrentPPO further improve stability and temporal awareness through distributional learning and recurrent architectures (Bühler *et al*., 2019; Schulman *et al*., 2017b).

Despite these strong theoretical foundations, applying DRL in financial markets remains challenging. Market data are inherently noisy, non-stationary and subject to regime shifts, which can lead to unstable convergence and sensitivity to hyperparameters (Velay *et al*., 2023). In addition, transaction costs, limited sample sizes and high cross-asset correlations may distort policy learning and reduce out-of-sample reliability. For this reason, comparing DRL agents with traditional benchmarks – such as equal-weight and minimum-variance portfolios – provides an essential reference for assessing whether adaptive methods genuinely outperform passive diversification strategies (Ledoit and Wolf, 2004).

Empirical studies suggest that while DRL models can capture short-term dependencies and non-linear interactions among asset returns, their performance is highly sensitive to market structure and liquidity conditions (Espiga-Fernández *et al*., 2024; Rezaei and Nezamabadi-Pour, 2025). These challenges are particularly pronounced in emerging and frontier markets, where thinner liquidity, stronger retail participation and frequent structural changes introduce additional instability. At the same time, such environments provide a valuable test bed for evaluating the robustness and generalization of modern DRL algorithms under realistic constraints.

Motivated by these considerations, this study addresses the need for a controlled and reproducible evaluation of DRL-based portfolio strategies in an emerging-market setting. Specifically, this paper conducts a unified comparison of seven DRL algorithms – A2C, PPO, DDPG, TD3, SAC, TQC and RecurrentPPO – applied to five Vietnamese equities over the period 2018–2025. By maintaining a consistent experimental pipeline, incorporating realistic transaction costs and performing multiple random-seed replications, the study aims to isolate algorithmic behavior from implementation-specific variations and provide transparent, reproducible evidence on model performance.

The main contributions of this study are threefold. First, this study develops a unified and fully controlled experimental framework for evaluating multiple DRL algorithms under identical market conditions, enabling a fair and transparent comparison across on-policy, off-policy, deterministic, stochastic and distributional approaches. Second, it provides empirical evidence on how key algorithmic design features – such as entropy regularization, clipped policy updates and recurrent state representations – affect return dynamics and risk-adjusted performance in an emerging and frontier market setting characterized by non-stationarity and limited liquidity. Third, the study ensures reproducibility and practical relevance by incorporating realistic transaction costs and conducting multiple independent training runs with different random seeds, thereby isolating algorithmic behavior from implementation-specific variations and offering robust and replicable insights into DRL-based portfolio performance. Compared with prior studies that typically focus on a single algorithm or heterogeneous experimental setups, this framework provides a more controlled and comparable evaluation environment, particularly suited for emerging-market conditions.

The selected equities – SBT.VN, BID.VN, CTG.VN, HPG.VN and VCB.VN – represent key sectors of the Vietnamese economy, including finance, manufacturing and consumer goods. These stocks are among the most actively traded and consistently listed constituents of the Ho Chi Minh Stock Exchange, offering relatively long and reliable data histories compared with other Vietnamese equities. This manageable yet diverse universe supports robust time-series modeling while preserving the complexity of a multi-asset environment. Moreover, differences in liquidity and market depth across these assets reflect real frontier-market conditions, making them suitable for testing the adaptability and stability of DRL-based portfolio strategies. The emphasis of this study is not on algorithmic novelty but on isolating the effect of design choices under a unified and controlled experimental pipeline.

The remainder of the paper is organized as follows: Section 2 reviews prior research on DRL in portfolio management, emphasizing risk sensitivity, stability and market adaptability. Section 3 describes the data, feature construction and benchmark portfolios. Section 4 presents the methodological framework, including the environment design, reward formulation and training–evaluation setup. Section 5 reports empirical results on convergence, comparative performance and robustness under different market conditions. Finally, Section 6 concludes and discusses implications for future research.

2. Literature review

RL has gained traction in financial portfolio management for its ability to learn optimal allocation strategies through repeated interaction with dynamic markets (Sutton and Barto, 2018). DRL extends this framework by integrating neural networks, enabling agents to approximate complex value functions and policies in high-dimensional, non-linear environments (Jiang *et al*., 2017; Deng *et al*., 2016). Recent developments such as A2C, PPO (Schulman *et al*., 2017b), DDPG, TD3 (Fujimoto *et al*., 2018) and SAC (Haarnoja *et al*., 2018) have established DRL as a viable approach for continuous-action portfolio optimization.

Although these advances are significant, most agents remain prone to overfitting historical patterns and performing poorly under regime shifts or structural breaks (Velay *et al*., 2023). To enhance robustness, researchers have proposed hybrid or risk-sensitive variants that penalize volatility and drawdowns (Rezaei and Nezamabadi-Pour, 2025). Entropy-based methods such as SAC encourage exploration and produce smoother learning, while actorcritic families (A2C, PPO) balance bias and variance in policy updates. Off-policy models (DDPG, TD3) improve sample efficiency through experience replay but often become unstable when market noise dominates. These issues motivate comparative studies evaluating multiple algorithms under identical market conditions and transaction-cost structures.

Recent literature also introduces advanced architectures such as TQC and RecurrentPPO to address heavy-tailed rewards and partial observability (Bühler *et al*., 2019; Schulman *et al*., 2017b). TQC extends SAC by truncating extreme critic quantiles, improving value-estimate stability in volatile markets, while RecurrentPPO employs long short-term memory (LSTM) units to preserve temporal memory, which is useful when markets exhibit serial dependence or delayed reactions. Empirical results show both families deliver more consistent out-of-sample performance, making them promising for multi-asset trading environments with strong autocorrelation and heteroscedasticity.

Benchmarking and risk management remain essential for financial validation. Conventional baselines such as equal-weight (EW) and minimum-variance (MinVar) portfolios provide clear yardsticks for evaluating DRL agents. MinVar typically relies on covariance-matrix estimation; the Ledoit–Wolf shrinkage estimator (Ledoit and Wolf, 2004) is preferred for high-dimensional assets because it yields well-conditioned, positive-definite matrices even with limited data. Comparing DRL agents against EW and MinVar benchmarks helps determine whether adaptive methods truly add value beyond passive diversification.

Another related strand of research combines price forecasting models with classical portfolio optimization, such as LSTM-based return prediction followed by Markowitz mean–variance allocation (Hochreiter and Schmidhuber, 1997; Markowitz, 1952). Such forecast-then-optimize approaches are intuitive and useful, especially when the main objective is to improve expected-return estimation. However, they separate the prediction stage from the allocation stage, which may lead to error propagation when forecast errors are passed into the optimizer. In contrast, DRL learns allocation policies directly from market interaction and can incorporate rebalancing behavior, transaction costs and dynamic risk–return trade-offs within a single decision-making framework (Jiang *et al*., 2017). In this sense, DRL differs from classical mean–variance allocation and covariance-based benchmarks, including those using shrinkage covariance estimators (Ledoit and Wolf, 2004), because it learns portfolio decisions

endogenously rather than solving a separate optimization problem from predicted returns. Therefore, the contribution of DRL is not only improved prediction accuracy but also the ability to learn adaptive portfolio decisions under changing market conditions.

Within emerging markets such as Vietnam, DRL applications are still limited but promising. Vietnamese equities feature moderate liquidity, sectoral concentration and heavy retail participation, producing frequent non-stationarity and volatility clustering. Studies on Vietnamese stocks indicate these characteristics create an ideal testbed for adaptive algorithms capable of responding to rapid structural changes (Saberironaghi *et al*., 2025). However, most prior work focuses on forecasting rather than allocation. A gap remains in systematically assessing multiple DRL algorithms under realistic constraints – transaction costs, frequent rebalancing and stochastic volatility – within reproducible setups. This study addresses that gap through a unified multi-agent backtest on five major Vietnamese equities (SBT.VN, BID.VN, CTG.VN, HPG.VN and VCB.VN) from 2018 to 2025, comparing seven DRL algorithms (A2C, PPO, DDPG, TD3, SAC, RecurrentPPO and TQC) against EW and MinVar benchmarks. By emphasizing reproducibility, consistency across ten random-seed runs and transparent evaluation metrics, this work contributes to methodological refinement in DRL portfolio research and empirical understanding of frontier-market dynamics.

Recent evidence further shows that Southeast Asian and Vietnamese equity markets remain structurally different from mature markets in terms of liquidity dynamics, investor composition and volatility behavior. The Association of Southeast Asian Nations (ASEAN) markets experienced pronounced liquidity changes around major shocks, while Vietnam's capital market continues to be characterized by strong retail participation and market-depth constraints (Thi Bich, 2025; World Bank, 2024). These structural features strengthen the case for adaptive DRL frameworks, especially recurrent architectures such as RecurrentPPO, which are designed to retain temporal information and respond to unstable market signals.

3. Data

3.1 Universe and sample period

The empirical universe comprises five Vietnamese equities actively traded on the Ho Chi Minh Stock Exchange (HOSE): SBT.VN, BID.VN, CTG.VN, HPG.VN and VCB.VN. These assets represent diverse industry exposures including finance, manufacturing and consumer sectors.

The five stocks were selected based on three criteria: active trading and data availability over the full sample period, sectoral diversification across finance, manufacturing and consumer-related industries and representation of large and liquid Vietnamese equities commonly associated with the VN30 investable universe. This selection ensures that the experiment uses assets with sufficient market depth and reliable price histories while maintaining a controlled portfolio dimension for cross-algorithm comparison.

The five equities were selected not only for their liquidity but also for their continuous data availability, sectoral representation and suitability for a controlled comparative experiment. Specifically, the sample includes major financial, industrial and consumer-related firms, allowing the agents to learn from assets with different return dynamics while maintaining a balanced and reliable panel. This relatively compact universe also helps reduce noise from thinly traded or discontinuously listed stocks, which is particularly important in emerging-market DRL applications.

The analysis is conducted on a focused set of highly liquid Vietnamese equities to ensure data consistency and model stability. While this design enables controlled experimentation, future work should extend the framework to larger and more diverse asset universes. This focused design is intentional to maintain a controlled experimental setting; however, future work should extend the framework to larger and more diverse asset universes to validate generalizability.

Daily open, high, low, close, volume (OHLCV) data are obtained through a local-first process: the system loads pre-downloaded comma-separated values (CSV) or Parquet files

when available and reverts to Yahoo Finance only when local data are missing. The full sample covers January 1, 2018, to September 30, 2025. The period from January 1, 2018 to January 1, 2023 is used for training, while January 1, 2023, to September 30, 2025, is reserved for out-of-sample evaluation. Trading calendars across assets are aligned to produce a balanced panel, with forward–backward filling applied for missing records due to market holidays.

3.2 Feature engineering

For each asset, a consistent set of technical indicators is computed using the ta-library:

(1) RSI(14): Relative strength index over a 14-day window, indicating overbought or oversold conditions;

(2) MACD and MACD_Signal: Moving average convergence–divergence and its signal line for trend-reversal detection;

(3) SMA(20) and EMA(20): Simple and exponential moving averages with a 20-day window to smooth short-term volatility;

(4) Bollinger bands (window 20, 2σ): capture relative volatility and price extremes and

(5) OBV: On-balance volume, representing directional trade-volume flow.

All indicators are forward- and backward-filled within each ticker to preserve temporal continuity. Dynamic covariance estimation via the Ledoit–Wolf shrinkage estimator (Ledoit and Wolf, 2004) is applied only to construct the MinVar benchmark and is excluded from the reinforcement learning state to isolate the contribution of technical features.

3.3 Benchmark portfolios

Two deterministic benchmarks are computed for comparison:

(1) Equal-weight (EW): simple average of daily returns across all assets.

(2) Minimum-variance (MinVar): daily rebalanced using the Ledoit–Wolf shrinkage covariance estimated from 252-day trailing returns underweight bounds (0, 0.1).

4. Methodology

4.1 Feature engineering

The feature set comprises eight technical indicators per asset: RSI(14), MACD, MACD_signal, SMA(20), EMA(20), Bollinger upper band, Bollinger lower band and OBV. The indicators are computed separately for each asset to avoid look-ahead bias and merged into a synchronized panel. Turbulence and covariance terms are excluded from the RL state, as the Ledoit–Wolf covariance estimator (Ledoit and Wolf, 2004) is applied only to the MinVar benchmark.

4.2 Environment formulation

A custom gymnasium environment, StockPortfolioEnv, is developed to simulate multi-asset portfolio trading with transaction costs and daily rebalancing through continuous actions. The environment is fully differentiable and compatible with continuous-control DRL algorithms such as PPO, SAC, TD3 and RecurrentPPO.

The environment follows the same general logic as widely used financial DRL benchmarking frameworks such as FinRL, where trading environments, agents and backtesting modules are organized to support reproducible portfolio-learning experiments (Liu *et al*., 2020, 2021). However, the present environment is deliberately kept compact and problem-specific, focusing on five Vietnamese equities, continuous portfolio weight

allocation, proportional transaction costs and a unified comparison of seven DRL agents under identical data and evaluation settings.

(1) *State space:* The environment state at time t is represented as

$$s_t = \left[V_t,\ p_t,\ h_t,\ \mathrm{RSI}_t,\ \mathrm{MACD}_t,\ \mathrm{MACD\_signal}_t,\ \mathrm{SMA}_{20,t},\ \mathrm{EMA}_{20,t},\ \mathrm{BB}_{\mathrm{high},t},\ \mathrm{BB}_{\mathrm{low},t},\ \mathrm{OBV}_t \right],$$

where V_t denotes current portfolio value, $p_t \in \mathbb{R}^N$ the vector of asset prices and $h_t \in \mathbb{R}^N$ current asset holdings. Technical indicators (RSI, MACD, SMA, EMA, Bollinger bands and OBV) provide temporal and trend-related information for each asset. No turbulence or covariance features are included in this study to isolate the impact of technical indicators.

(2) *Action space:* Each agent outputs a continuous vector $a_t \in [-1, 1]^N$, which is mapped to portfolio weights w_t via the softmax normalization:

$$w_{i,t} = \frac{\exp(a_{i,t})}{\sum_{j=1}^{N} \exp(a_{j,t})}, \quad \text{for } i = 1, \ldots, N.$$

This transformation guarantees $w_{i,t} \geq 0$ and $\sum_i w_{i,t} = 1$, ensuring a long-only and fully invested portfolio. Therefore, the agent continuously adjusts allocation proportions across assets rather than issuing discrete *buy/hold/sell* commands.

(3) *Transaction costs:* A fixed cost of 10 basis points (0.1%) is applied to each buy or sell operation, calculated on the traded notional value.

(4) *Portfolio dynamics:* After executing an action at time t, the portfolio value evolves as

$$V_{t+1} = \mathrm{cash}_t - \sum_i \Delta v_{i,t} - C_t + \sum_i \frac{\widehat{v}_{i,t}}{p_{i,t}}\, p_{i,t+1},$$

where $\Delta v_{i,t}$ represents the net value traded in asset i and C_t the total transaction cost.

(5) *Reward:* The immediate reward is defined as the normalized daily portfolio return:

$$r_{t+1} = \frac{V_{t+1} - V_t}{\max(V_t, \varepsilon)} \times 10^{-4},$$

which is implemented without additional penalty terms or risk adjustments to preserve comparability across agents. This design isolates return-driven learning and ensures that differences in performance are attributable to algorithmic characteristics rather than reward engineering.

The use of a return-based reward is an intentional design choice. By avoiding additional risk-adjustment terms in the reward function, the experiment preserves a controlled comparison across algorithms and isolates the effect of algorithmic design rather than reward engineering. However, this simplified reward specification may also encourage some agents to pursue higher-return but higher-volatility strategies, since drawdowns and volatility are evaluated ex post rather than penalized directly during training.

All agents thus operate under a continuous control policy, learning to rebalance allocation weights smoothly at each step while accounting for transaction costs and market feedback. To further interpret this setup, the use of continuous action spaces enables fine-grained portfolio adjustments and smoother rebalancing behavior, which are particularly important in financial markets with transaction costs and rapidly changing conditions. This design choice

supports more realistic trading behavior compared to discrete allocation schemes, where abrupt changes in portfolio weights may lead to excessive turnover.

In addition to proportional transaction costs, real-world trading in emerging markets may also be affected by liquidity constraints, slippage, etc. These frictions are not explicitly modeled in the present environment, which it assumes execution at observed prices plus a fixed transaction cost. Therefore, the reported results can be interpreted as conservative evidence under simplified execution assumptions rather than as a full representation of real trading impact.

All algorithms in this study operate within a *continuous action space*, suitable for fine-grained portfolio rebalancing rather than discrete buy/hold/sell actions. Specifically, A2C, PPO, SAC, DDPG, TD3, TQC and RecurrentPPO output a real-valued action vector representing the target allocation weights across N assets. These actions are first normalized via a softmax transformation to ensure the portfolio remains fully invested and long-only. Hence, each element of the action vector determines the proportion of wealth allocated to a specific stock rather than discrete trading decisions. Although the environment internally computes transaction costs for both increasing ("buy") and decreasing ("sell") positions, the agent does not explicitly select between discrete actions such as buy, hold or sell; instead, it continuously adjusts portfolio weights at each step to optimize long-term return.

In summary, all evaluated DRL algorithms use continuous control policies: on-policy agents (A2C, PPO and RecurrentPPO) learn smooth stochastic policies parameterized by Gaussian distributions, while off-policy agents (DDPG, TD3, SAC and TQC) employ deterministic or entropy-regularized continuous actions for high-resolution rebalancing in continuous time.

4.3 Reinforcement learning algorithms

Seven DRL algorithms are evaluated: A2C, PPO, SAC, DDPG, TD3, TQC and RecurrentPPO. All are implemented in Stable-Baselines3, with sb3_contrib extensions where available. A2C and PPO are on-policy actor–critic models, while DDPG, TD3 and SAC are off-policy methods using replay buffers and target networks. TQC and RecurrentPPO extend these designs by improving value distribution accuracy and capturing temporal dependencies. Each algorithm operates in continuous action spaces and adapts to volatile financial conditions (Espiga-Fernández *et al*., 2024; Liu *et al*., 2022; Rezaei and Nezamabadi-Pour, 2025; Bühler *et al*., 2019; Schulman *et al*., 2017b).

These algorithms are selected to represent a diverse set of learning paradigms, including on-policy and off-policy training and deterministic and stochastic policies, as well as distributional and recurrent extensions.

This diversity enables a systematic and controlled comparison of learning stability, exploration behavior and robustness under non-stationary and noisy financial environments.

All agents are trained on the in-sample window for 150,000 timesteps using a vectorized environment (DummyVecEnv) to parallelize rollouts. Random seeds are fixed for Python, NumPy and PyTorch to ensure reproducibility.

To improve reproducibility, Table 1 summarizes the main hyperparameter settings used for each DRL algorithm. These parameters were kept fixed across all experimental replications, except for the random seeds, to ensure that performance differences mainly reflect algorithmic behavior rather than tuning variation. All models were implemented using the same Stable-Baselines3-based pipeline, identical training and testing windows and the same transaction-cost setting.

Although the hyperparameters were not exhaustively optimized, they were selected from commonly used Stable-Baselines3 configurations and kept consistent across runs to preserve a fair comparative design.

4.4 Algorithmic foundations

Each DRL agent follows the actor–critic paradigm with distinctive update objectives and stability mechanisms designed for continuous control.

Table 1. Key hyperparameters for DRL algorithms

Algorithm	Learning rate	n_steps/batch	Buffer size	Notes
A2C	1×10^{-4}	10/–	–	GAE 0.95
PPO	5×10^{-4}	1,024/64	–	entropy coef 0.01
SAC	3×10^{-4}	–/256	10^6	ent_coef auto 0.2, log std −3
DDPG	3×10^{-4}	–/256	10^6	noise $\sigma = 0.1$
TD3	1×10^{-3}	–/256	10^6	noise $\sigma = 0.1$
TQC	3×10^{-4}	–/512	10^6	ent_coef auto 0.1, quantile trunc. 25%
RecurrentPPO	3×10^{-4}	1,024/128	–	LSTM size 128, 1 layer

Note(s): GAE: generalized advantage estimation

The inclusion of multiple algorithmic variants allows a systematic comparison of learning stability, exploration behavior and robustness under non-stationary financial environments.

4.4.1 Advantage actor–critic (A2C).

$$L_{A2C} = \frac{1}{N}\sum_{i=1}^{N}[\log\pi_\theta(a_i|\ s_i)A_i] - c\,\frac{1}{N}\sum_{i=1}^{N}\left(V(s_i) - R_i\right)^2. \tag{1}$$

A2C estimates both the policy gradient and the state-value function synchronously, providing simplicity and stability but limited sample efficiency under non-stationary market conditions (Mnih *et al*., 2016).

4.4.2 Proximal policy optimization (PPO).

$$L_{PPO}(\theta) = \frac{1}{N}\sum_{i=1}^{N}\min(r_i(\theta)A_i, \text{clip}(r_i(\theta), 1-\epsilon, 1+\epsilon)A_i), \tag{2}$$

where $r_i(\theta) = \frac{\pi_\theta(a_i|s_i)}{\pi_{\theta_{\text{old}}}(a_i|s_i)}$. The clipping mechanism constrains policy updates to prevent large deviations, improving training stability and robustness in continuous-action environments (Schulman *et al*., 2017a,b).

4.4.3 Deep deterministic policy gradient (DDPG).

$$L_{DDPG} = \frac{1}{N}\sum_{i=1}^{N}\left(y_i - Q\left(s_i, a_i|\ \theta^Q\right)\right)^2, \quad y_i = r_i + \gamma Q'\left(s_{i+1}, \mu'\left(s_{i+1}|\ \theta^{\mu'}\right)\right). \tag{3}$$

DDPG combines deterministic policy gradients with replay-buffer experience reuse, enabling efficient learning in high-dimensional continuous spaces but susceptible to overestimation bias and instability under noise (Lillicrap *et al*., 2015; Fujimoto *et al*., 2018).

4.4.4 Twin delayed DDPG (TD3).

$$Q_{\text{target}} = r + \gamma\min\left(Q_{\theta'_1}(s', a'), Q_{\theta'_2}(s', a')\right), \tag{4}$$

reducing DDPG's overestimation bias via twin critics and delayed updates (Rezaei and Nezamabadi-Pour, 2025).

4.4.5 Soft actor–critic (SAC).

$$L_{\text{actor}}(\phi) = \frac{1}{N}\sum_{i=1}^{N}\mathbb{E}_{a_i\sim\pi_\phi}\left[\alpha\log\pi_\phi(a_i|\ s_i) - Q_{\theta^Q}(s_i, a_i)\right]. \tag{5}$$

Entropy regularization promotes exploration and robustness to regime shifts (Rezaei and Nezamabadi-Pour, 2025).

4.4.6 Truncated quantile critics (TQC). TQC extends SAC by truncating upper quantiles of the value distribution to prevent overestimation (Bühler *et al*., 2019):

$$L_{TQC} = \frac{1}{N}\sum_{i=1}^{N} \|\text{Huber}\left(Z_\theta(s_i, a_i) - TZ_{\bar{\theta}}(s_{i+1}, a_{i+1})\right)\|_2. \quad (6)$$

4.4.7 Recurrent proximal policy optimization (RecurrentPPO). Incorporates an LSTM layer into PPO to capture temporal dependencies (Schulman *et al*., 2017b):

$$h_t = f_\theta(s_t, h_{t-1}), \quad \pi_\theta(a_t|\, s_t, h_{t-1}),\ V_\theta(s_t, h_{t-1}),$$

enhancing continuity in time-correlated markets.

4.5 Training and evaluation protocol

Each agent trains for 150,000 timesteps and is evaluated deterministically on the out-of-sample period. Training and testing repeat ten times with seeds $42 + r, r = 1,\ldots, 10$, to average stochastic variability. Performance metrics include cumulative and annualized return, annualized volatility, Sharpe ratio and maximum drawdown. All results are logged and aggregated across runs.

4.6 Benchmarks

Two benchmark strategies are used:

(1) Equal-weight portfolio: average of daily returns across all assets and

(2) Minimum-variance portfolio: daily rebalanced using PyPortfolioOpt with 252-day rolling covariance and Ledoit–Wolf shrinkage (Ledoit and Wolf, 2004).

While a buy-and-hold VN-Index provides a reference for overall market growth, this study focuses on cross-sectional portfolio allocation within a fixed asset universe. Therefore, equal-weight and minimum-variance portfolios are used as benchmarks to ensure a consistent comparison with DRL agents under identical constraints.

4.7 Performance evaluation metrics

Model performance is assessed using standard measures:

(1) Cumulative return: $\prod_{t=1}^{T}(1 + r_t) - 1$

(2) Annualized return: $(1 + \text{Cumulative Return})^{252/T} - 1$

(3) Annualized volatility: $\sqrt{252}\,\sigma(r_t)$

(4) Sharpe ratio: $\frac{\mathbb{E}[r_t - r_f]}{\sigma(r_t)}$ (assume $r_f = 0$)

(5) Maximum drawdown: $\max_{t\in[0,T]}\left(\frac{\max_{\tau\le t} V_\tau - V_t}{\max_{\tau\le t} V_\tau}\right)$

4.8 Convergence analysis protocol

Each model is checkpointed every 25,000 steps and evaluated on the trading window. Policies are reloaded without retraining, and out-of-sample performance is measured on the same metrics to visualize convergence for Sharpe ratio, return, volatility and drawdown.

4.9 Summary

The workflow consists of five stages: (1) data preprocessing and feature extraction; (2) environment setup with rebalancing and transaction costs; (3) DRL training (A2C, PPO,

DDPG, TD3, SAC, TQC and RecurrentPPO); (4) benchmark computation and (5) out-of-sample evaluation and convergence analysis.

5. Results

This section presents the empirical results of the DRL portfolio experiments conducted on five Vietnamese equities (SBT.VN, BID.VN, CTG.VN, HPG.VN and VCB.VN). All models were trained on daily data from 2018 to 2024 and evaluated over the 2023–2025 trading horizon using ten random-seed replications. The analysis focuses on cumulative returns, convergence stability, risk–return trade-offs and comparative robustness against traditional benchmarks. Overall, DRL-based strategies tend to outperform equal-weight and minimum-variance baselines on most evaluation metrics, with a few noted exceptions, indicating superior adaptability under emerging-market conditions.

5.1 Cumulative return dynamics

Figure 1 shows the median cumulative returns of the DRL agents – A2C, PPO, DDPG, TD3, SAC, TQC and RecurrentPPO – together with the equal-weight and minimum-variance benchmarks. All agents achieved positive cumulative growth, with divergence becoming more visible after mid-2024. Among them, PPO exhibits the strongest and most persistent upward trajectory, clearly outperforming all other agents and benchmarks throughout most of the evaluation horizon. TD3 ranks second, displaying comparable growth with slightly higher short-term fluctuations, reflecting its deterministic yet responsive exploration strategy. A2C remains tightly coupled to the equal-weight baseline, suggesting limited flexibility under regime changes. Overall, the results confirm PPO's superior ability to capture persistent market trends and maintain convergence stability, with TD3 emerging as a solid but more volatile alternative.

5.2 Convergence stability

Figure 2 illustrates the checkpoint-based convergence of the Sharpe ratio for a representative training run. Each colored line corresponds to an agent in Figure 2: PPO (green), TD3 (cyan), RecurrentPPO (light purple), A2C (blue-violet), DDPG (orange), SAC (light orange) and TQC (pink). Across 25,000–150,000 training steps, most algorithms display stable and gradually improving Sharpe ratios, confirming consistent policy learning and numerical reliability.

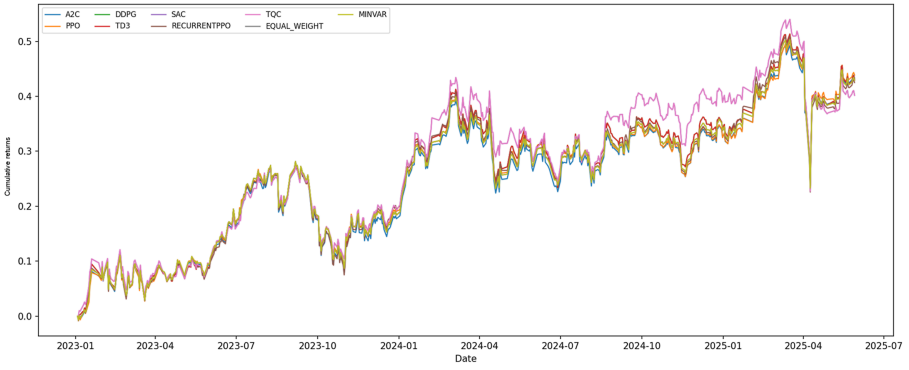

Figure 1. Median cumulative returns of DRL agents and benchmarks (2023–2025). PPO achieves the highest cumulative performance, followed by TD3, while SAC and TQC remain closer to benchmark levels. The results represent the median of ten independent runs using five Vietnamese equities

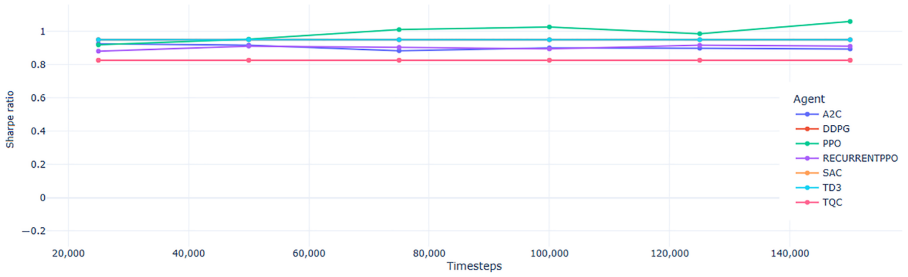

Figure 2. Checkpoint-based Sharpe ratio convergence for representative training runs. PPO (green) achieves the highest and most consistent improvement, followed by TD3 (cyan) and RecurrentPPO (light purple), whereas SAC (light orange) and TQC (pink) remain comparatively flat

PPO (green) exhibits the most pronounced upward trajectory, surpassing all others and steadily improving beyond a Sharpe of 1.0 by the end of training. TD3 (cyan) and RecurrentPPO (light purple) follow closely, maintaining smooth convergence with moderate variance. A2C (blue-violet) stabilizes early near 0.9, indicating limited exploration capacity under synchronous updates. In contrast, SAC (light orange) and TQC (pink) show nearly flat curves around 0.85–0.9, suggesting conservative learning dynamics and slower adaptation to changing reward structures. Overall, these convergence patterns reaffirm the stability advantage of PPO and its clipped-surrogate policy design, while TD3 and RecurrentPPO provide robust but slightly less responsive alternatives.

5.3 Risk–return and risk-adjusted performance

Figure 3 compares the median annual return (blue) and median maximum drawdown (red) across ten replications. All DRL agents achieve annualized returns between 15 and 17% on average, exceeding the benchmarks. PPO delivers the highest annual return (0.1666) and the best downside profile by Calmar (0.9606) and Sortino (1.3120), indicating that it captures trends while containing drawdowns. RecurrentPPO trades some return (0.1618) for superior risk adjustment, achieving the top Sharpe ratio (1.0461). By contrast, SAC and TQC earn lower total returns (0.1537 each) and weaker drawdown efficiency (Calmar 0.7866) despite Sharpe ratios just above 1.0; their Omega (1.1468) and Sortino (1.1343) also trail PPO and even sit close to or below the equal-weight/min-var baselines (Omega 1.1792, Sortino 1.2570). TD3 and DDPG are mid-pack (returns 0.1638; Sharpe 0.8842), while A2C aligns with the baselines on Sharpe (0.8976) but offers no advantage in drawdown control.

To complement the figure, Table 2 reports median annualized return, volatility, Sharpe, Calmar, Omega, Sortino and training stability. Three points stand out from the table: (1) *PPO* combines the top return (0.1666) with the strongest Calmar (0.9606) and Sortino (1.3120), giving the best return–drawdown trade-off; (2) RecurrentPPO yields the highest Sharpe (1.0461), confirming its risk-adjusted edge despite a slightly lower return; (3) SAC/TQC, while clearing Sharpe $\approx$ 1.04, underperform on total return and drawdown efficiency (Calmar 0.7866; Omega 1.1468), so they do not dominate the benchmarks. Overall, PPO and RecurrentPPO provide the most attractive balance of profitability and risk control, whereas SAC/TQC are more conservative and less efficient on downside metrics in this setting.

It is worth noting that some agents with relatively high annual returns nonetheless record lower Sharpe ratios. This occurs because Sharpe ratio measures excess return per unit of volatility. Thus, when a model like DDPG or A2C generates higher raw returns accompanied by proportionally higher variance or more frequent drawdowns, the resulting Sharpe ratio declines even if the mean return is strong. In other words, high annual returns in these models stem from aggressive allocation shifts and momentum chasing during short-term rallies, which amplify gains but also increase downside exposure. PPO and RecurrentPPO, by contrast,

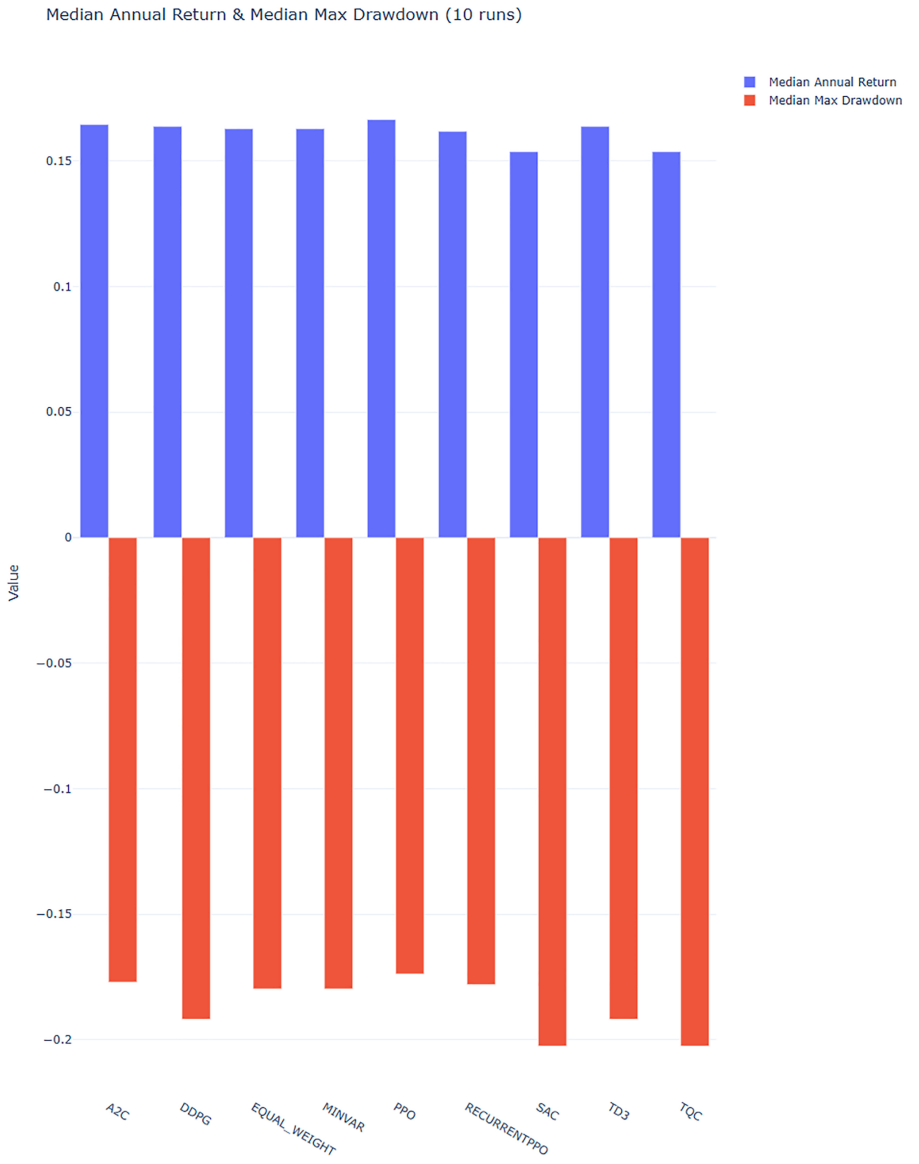

Figure 3. Median annual return and maximum drawdown of DRL agents and benchmarks over ten runs. Blue bars indicate return; red bars indicate drawdown

maintain smoother volatility profiles and smaller tail losses, yielding more stable risk-adjusted outcomes even if their nominal returns are only marginally higher. This distinction illustrates that profitability alone is not sufficient; risk normalization via Sharpe or Calmar ratios provides a more accurate gauge of long-term portfolio efficiency.

In terms of Sharpe ratio leadership, RecurrentPPO (1.0461) stands out as the top performer, followed closely by SAC (1.0406) and TQC (1.0406). PPO, while slightly lower on Sharpe (0.9251), remains the most balanced model overall due to its combination of higher return, superior Calmar and stronger Sortino ratio. This pattern suggests that the Sharpe-optimal

Table 2. Median performance metrics of DRL agents and benchmarks (2018–2025)

Model	Ann. Return	Ann. Vol.	Sharpe	Calmar	Omega	Sortino	Stability
A2C	0.1646	0.1879	0.8976	0.9232	1.1790	1.3092	0.7893
DDPG	0.1638	0.1923	0.8842	0.8148	1.1767	1.2432	0.7793
EQUAL_WEIGHT	0.1629	0.1862	0.9017	0.9064	1.1792	1.2570	0.7948
MINVAR	0.1629	0.1862	0.9017	0.9064	1.1792	1.2570	0.7948
PPO	*0.1666*	0.1894	0.9251	*0.9606*	*1.1820*	*1.3120*	*0.7856*
RecurrentPPO	0.1618	0.1929	*1.0461*	0.8914	1.1539	1.1188	0.7390
SAC	0.1537	0.1913	1.0406	0.7866	1.1468	1.1343	0.7283
TD3	0.1638	0.1923	0.8842	0.8148	1.1767	1.2432	0.7793
TQC	0.1537	0.1913	1.0406	0.7866	1.1468	1.1343	0.7283

agents (RecurrentPPO, SAC and TQC) achieve steadier risk-adjusted growth, whereas PPO achieves the highest absolute gain without significant volatility penalties – an important trade-off between stability and profitability in emerging-market contexts.

Figure 4 displays the distribution of Sharpe ratios across the ten experimental runs. PPO attains the highest median Sharpe ratio (around 0.93–0.94) with the tightest interquartile range, indicating both strong and consistent risk-adjusted performance across runs. A2C performs comparably on median value but exhibits slightly wider dispersion, suggesting moderate volatility sensitivity. RecurrentPPO, SAC, TD3 and TQC show greater variability and lower medians (around 0.85–0.88), implying less consistent out-of-sample efficiency. The baseline strategies (equal-weight and min-variance) maintain stable but lower Sharpe levels near 0.90, confirming their conservative nature. Overall, PPO emerges as the most balanced and reliable performer, combining the highest median Sharpe with the narrowest variability, reflecting robust policy convergence and stable generalization.

5.4 Summary of findings

Across analyses, PPO is the clear overall leader: it delivers the highest annual return (0.1666) and the strongest downside profile by Calmar (0.9606) and Sortino (1.3120), and it shows the most persistent cumulative-growth trajectory in Figure 1 alongside the smoothest Sharpe-ratio convergence in Figure 2. TD3 ranks second on cumulative growth but with slightly higher short-term fluctuations. RecurrentPPO provides the best risk-adjusted balance in the tabulated median (Sharpe = 1.0461) and maintains smooth learning, though it trades off some total return (0.1618). SAC and TQC register Sharpe values just above 1.0 in the table but underperform on total return (0.1537) and drawdown efficiency (Calmar 0.7866; Sortino 1.1343), leaving them close to or below the EW/MinVar baselines on several downside metrics. A2C and DDPG yield moderate results – with A2C tracking the baselines on Sharpe and DDPG/TD3 mid-pack – yet they offer less adaptability. Taken together, PPO offers the best all-around profile (return, stability and dispersion control); RecurrentPPO excels on risk-adjusted efficiency; TD3 is a solid but more volatile alternative and SAC/TQC behave more conservatively without dominating the benchmarks in this setting.

5.5 Discussion

The empirical results offer several insights into how DRL algorithms behave under the structural and liquidity constraints of the Vietnamese equity market. Although all models generated positive cumulative returns, the differences in performance reveal that algorithmic design choices – particularly entropy regularization, recurrent state encoding and clipped policy updates – determine the balance between profitability and stability. These findings should be interpreted within a controlled experimental setting, where differences in outcomes are primarily attributable to algorithmic design rather than variations in data preprocessing, feature construction or implementation details.

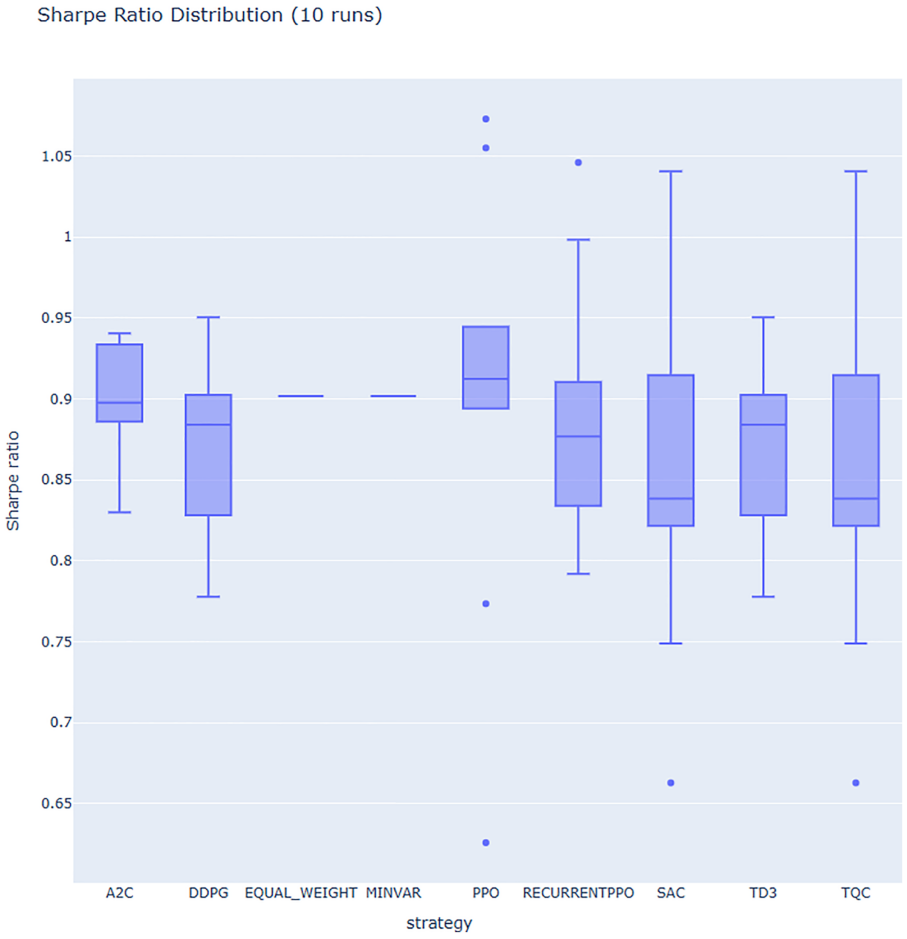

Figure 4. Distribution of Sharpe ratios across ten independent experimental runs. PPO attains the highest median and lowest dispersion, while other agents such as RecurrentPPO, SAC and TD3 show wider variability

First, PPO shows clear practical advantages as a stable and high-performing on-policy method for portfolio allocation. It achieved the highest annual return (0.1666) and the strongest Calmar (0.9606) and Sortino (1.3120) ratios, illustrating its ability to generate consistent profits with limited downside exposure. Its clipped surrogate objective constrains policy updates and prevents excessive parameter shifts that could destabilize learning under noisy conditions. This stability explains PPO's smooth convergence and moderate volatility despite market disruptions, consistent with the evidence of Schulman *et al*. (2017a) on the algorithm's steady performance in continuous-control settings. Importantly, Figure 4 also shows PPO with the highest median Sharpe (about 0.93–0.94) and the tightest dispersion across runs, corroborating its reliability. From a practical perspective, this suggests that PPO's policy update mechanism is well-suited for environments characterized by noisy signals and frequent regime shifts, where overly aggressive updates can degrade performance.

This finding is consistent with prior studies that highlight PPO's stability in continuous control tasks but contrasts with some financial applications where entropy-based methods dominate, suggesting that market structure plays a critical role.

Second, RecurrentPPO achieved the highest median Sharpe in the tabulated summary (1.0461) and demonstrated smooth convergence across checkpoints. Its LSTM-based memory allows the policy to retain temporal dependencies, which enhances learning stability in partially observable environments (Schulman *et al*., 2017b). This feature appears particularly valuable in Vietnam's market context, where autocorrelation and momentum effects often coincide with rapid structural changes. However, its more conservative learning dynamics slightly reduced raw returns compared to PPO, indicating a deliberate trade-off between responsiveness and risk control. The apparent difference between Figure 4 (where PPO's boxplot median is highest) and the table (where RecurrentPPO's median Sharpe is highest) reflects two complementary summaries: the boxplot shows distributional behavior on the 10-run panel, while the table reports the aggregated median statistic across replications. This highlights the importance of evaluating DRL models using both distributional and aggregated metrics, as reliance on a single summary statistic may obscure important aspects of performance variability.

Third, SAC and TQC delivered risk-adjusted scores just above 1.0 but posted lower total returns (0.1537) and weaker downside control (Calmar 0.7866; Sortino 1.1343). Their entropy-regularized and distributional critics improve exploration robustness (Haarnoja *et al*., 2018; Bühler *et al*., 2019); yet, in this dataset, they did not translate that robustness into superior return–drawdown efficiency relative to PPO or even EW/MinVar on some downside metrics. TQC's quantile truncation still mitigates overestimation, but the overall profile remains conservative. This outcome suggests that exploration-oriented mechanisms, while theoretically appealing, may lead to overly cautious policies in relatively small or noisy datasets, limiting their ability to exploit profitable opportunities. This contrasts with prior findings where entropy-based methods often outperform in larger or more liquid markets, suggesting that their advantages may diminish under data constraints and market inefficiencies.

Fourth, A2C and DDPG displayed moderate to weaker risk-adjusted outcomes relative to other agents. A2C's synchronous updates limited adaptability to regime changes, while DDPG's deterministic policies were more sensitive to reward noise during turbulent phases (Fujimoto *et al*., 2018). Nevertheless, both maintained average returns above traditional benchmarks, indicating that even baseline DRL architectures can outperform static portfolios under emerging-market dynamics (Rezaei and Nezamabadi-Pour, 2025). This reinforces the broader insight that adaptive policy learning, even without advanced stabilization mechanisms, can capture structural inefficiencies in less efficient markets.

Finally, benchmarking against equal-weight and minimum-variance portfolios reinforces the effectiveness of the DRL framework. The MinVar benchmark – based on the Ledoit–Wolf shrinkage estimator (Ledoit and Wolf, 2004) – provides stability but lacks adaptability. In contrast, PPO, TD3 and RecurrentPPO adaptively rebalance exposure and learn dynamic risk management, producing higher cumulative growth and better multi-metric balance; SAC/TQC remain more defensive and do not dominate the baselines on drawdown-sensitive measures. This comparison underscores the importance of adaptability in portfolio allocation, particularly in environments where return distributions and correlations evolve over time.

While multiple random seeds are used to ensure robustness across training runs, this study does not perform an extensive sensitivity analysis on hyperparameters or transaction cost assumptions. Future research should examine the extent to which these factors influence performance stability.

A further practical consideration concerns execution quality. Although the environment incorporates proportional transaction costs, it does not explicitly model slippage or liquidity constraints. This omission may matter in the Vietnamese market, where order-book depth can vary across assets and trading sessions. In practice, larger rebalancing trades may be executed at less favorable prices than those observed in daily data, which could reduce realized returns and increase turnover-related losses. Accordingly, the out-of-sample results should be

interpreted as reflecting a controlled backtesting framework rather than full market-impact-adjusted execution.

This study does not include a VN-Index benchmark, as the objective is to evaluate allocation strategies within a fixed asset set. Equal-weight and minimum-variance portfolios therefore provide more appropriate comparators by isolating allocation effects rather than market-wide movements.

In summary, DRL-based portfolio allocation can outperform static benchmarks in emerging markets when enhanced with mechanisms that promote exploration stability and temporal awareness. PPO emerges as the most balanced and practically robust algorithm, combining high profitability, steady Sharpe improvements and consistent convergence. RecurrentPPO complements this by offering the strongest risk-adjusted median in the tabulated results, while TD3 provides a responsive but slightly more volatile alternative. SAC and TQC contribute robustness in principle but, in this application, remain conservative and less efficient on downside control. Future studies should extend the evaluation horizon, incorporate transaction-cost and slippage effects and examine cross-market transferability to assess whether these advantages persist across economic contexts. More broadly, the findings highlight the value of controlled and reproducible evaluation frameworks for understanding the interaction between algorithmic design and market structure in DRL-based portfolio management.

6. Conclusion

This study investigates the application of DRL for multi-asset portfolio optimization in the Vietnamese stock market. Five representative equities – SBT.VN, BID.VN, CTG.VN, HPG.VN and VCB.VN – were selected from the Ho Chi Minh Stock Exchange for their liquidity, sectoral diversity and reliable data coverage. Seven DRL agents – A2C, PPO, DDPG, TD3, SAC, TQC and RecurrentPPO – were trained on daily price data from 2018 to 2024 and evaluated out-of-sample from 2023 to 2025, ensuring both stable and volatile market regimes were represented.

Methodologically, the framework compared stochastic and deterministic actor–critic structures to evaluate how exploration design affects portfolio stability. PPO proved the most balanced algorithm, achieving the highest annual return (0.1666) and superior Calmar and Sortino ratios, sustaining growth under volatility; it also showed the highest median Sharpe and lowest dispersion in the distributional analysis. RecurrentPPO attained the top Sharpe in the tabulated median (1.0461) via LSTM-based temporal awareness, providing consistent learning and risk control while accepting slightly lower returns. TD3 delivered the second-strongest cumulative growth but with more short-term variability. SAC and TQC achieved Sharpe just above 1.0 yet underperformed on total return and drawdown efficiency, reflecting a more conservative posture. Deterministic models (DDPG, TD3) were moderate but steady, while A2C was stable with limited adaptability. These results support arguments that entropy-driven objectives (Haarnoja *et al*., 2018) and recurrent encoding (Schulman *et al*., 2017b) enhance robustness in high-volatility, partially observable environments.

Empirically, PPO, RecurrentPPO and TD3 outperform equal-weight and minimum-variance benchmarks on key dimensions; SAC/TQC remains competitive on some risk-adjusted metrics but do not dominate the baselines on drawdown-sensitive measures. The ability of these DRL agents – especially PPO – to internalize dynamic risk management without relying on static covariance assumptions (Ledoit and Wolf, 2004) demonstrates the adaptability of learning-based frameworks to frontier markets. Overall, DRL-based systems represent a viable, data-driven alternative to classical optimization by learning to balance return and risk directly from market interactions.

A limitation of this study is the relatively small asset universe. Although the five-stock setting enables a controlled comparison of DRL algorithms, it does not represent a fully diversified portfolio. The limited number of assets may restrict diversification opportunities and affect the ability of the learned policies to generalize to larger, more heterogeneous

investment universes. Future research should extend the framework to broader Vietnamese and ASEAN equity portfolios to test whether the relative ranking of DRL algorithms remains stable.

Another limitation concerns the simplified reward function. Because the agents are trained using daily portfolio returns without explicit volatility or drawdown penalties, some models may adopt more aggressive allocation policies. Future research should examine whether risk-adjusted reward formulations, such as the differential Sharpe ratio, Sortino-based rewards or drawdown-aware penalties, improve downside control and reduce excessive risk-taking in emerging-market portfolio allocation.

Although the present analysis identifies PPO and RecurrentPPO as the leading models, a limitation is that the current results provide limited direct interpretability of their allocation mechanisms. In particular, the study does not fully decompose whether superior performance arises from dynamic rebalancing across assets or from concentration in one or two high-performing stocks. Future research should therefore include portfolio weight dynamics, asset-level attribution and turnover decomposition to better explain why specific DRL agents outperform others.

A further limitation concerns portfolio scalability. The five-equity universe was deliberately selected to create a controlled and reproducible benchmark in which algorithmic behavior can be isolated from excessive asset-level noise and missing-data problems. Nevertheless, the results should not be interpreted as direct evidence of scalability to larger universes such as the VN30 index. When the number of assets increases, the state and action spaces expand rapidly, and the set of feasible portfolio-weight combinations becomes substantially more complex. This curse of dimensionality can reduce sample efficiency, slow convergence, amplify estimation noise and weaken out-of-sample stability. Accordingly, the present results establish a controlled baseline for DRL-based portfolio allocation in Vietnam, while future research should examine whether the relative advantages of PPO, RecurrentPPO and TD3 persist in larger and more diversified portfolios.

6.1 Suggestions for future research

Future research should extend the analysis to longer horizons and alternative reward-shaping schemes emphasizing downside protection and recovery dynamics. Incorporating macroeconomic indicators, sentiment signals, or cross-asset correlations could enhance responsiveness to structural shifts. Future work could also incorporate market-level benchmarks (e.g. VN-Index) to evaluate performance relative to overall market trends. Finally, exploring ensemble DRL models, transfer learning across ASEAN markets and more realistic execution assumptions – including slippage, bid–ask spread dynamics and market-impact effects – would strengthen the robustness and generalizability of DRL applications in real-world portfolio management.

Future research should also compare DRL agents with forecast-then-optimize strategies, such as LSTM forecasting combined with Markowitz optimization, to further evaluate whether direct policy learning offers robust advantages over sequential prediction-based portfolio construction.

References

Bühler, H., Gonon, L., Teichmann, J. and Wood, B. (2019), "Deep hedging", *Quantitative Finance*, Vol. 19 No. 8, pp. 1271-1291, doi: 10.1080/14697688.2019.1571683.

Deng, Y., Bao, F., Kong, Y., Ren, Z. and Dai, Q. (2016), "Deep direct reinforcement learning for financial signal representation and trading", *IEEE Transactions on Neural Networks and Learning Systems*, Vol. 28 No. 3, pp. 653-664, doi: 10.1109/TNNLS.2016.2522401.

Espiga-Fernández, F., García-Sánchez, Á. and Ordieres-Meré, J. (2024), "A systematic approach to portfolio optimization: a comparative study of reinforcement learning agents, market signals, and investment horizons", *Algorithms*, Vol. 17 No. 12, p. 570, doi: 10.3390/a17120570.

Fujimoto, S., van Hoof, H. and Meger, D. (2018), "Addressing function approximation error in actor-critic methods", *arXiv preprint*, available at: https://arxiv.org/abs/1802.09477

Haarnoja, T., Zhou, A., Abbeel, P. and Levine, S. (2018), "Soft actor–critic: off-policy maximum entropy deep reinforcement learning with a stochastic actor", *arXiv preprint*, available at: https://arxiv.org/abs/1801.01290

Hochreiter, S. and Schmidhuber, J. (1997), "Long short-term memory", *Neural Computation*, Vol. 9 No. 8, pp. 1735-1780, doi: 10.1162/neco.1997.9.8.1735.

Jiang, Z., Xu, D. and Liang, J. (2017), "A deep reinforcement learning framework for the financial portfolio management problem", *arXiv preprint*, available at: https://arxiv.org/abs/1706.10059

Ledoit, O. and Wolf, M. (2004), "A well-conditioned estimator for large-dimensional covariance matrices", *Journal of Multivariate Analysis*, Vol. 88 No. 2, pp. 365-411, doi: 10.1016/S0047-259X(03)00096-4.

Lillicrap, T.P., Hunt, J.J., Pritzel, A., Heess, N., Erez, T., Tassa, Y., Silver, D. and Wierstra, D. (2015), "Continuous control with deep reinforcement learning", *arXiv preprint arXiv:1509.02971*.

Liu, X.-Y., Yang, H., Chen, Q., Zhang, R., Yang, L., Xiao, B. and Wang, C.D. (2020), "FinRL: a deep reinforcement learning library for automated stock trading in quantitative finance", arXiv preprint arXiv:2011.09607.

Liu, X.-Y., Yang, H., Gao, J. and Wang, C. (2021), "FinRL: deep reinforcement learning framework to automate trading in quantitative finance", 4 November, available at: https://ssrn.com/abstract=3955949, doi: 10.2139/ssrn.3955949.

Liu, X.-Y., Yang, H., Chen, Q., Zhang, R., Yang, L., Xiao, B. and Wang, C.D. (2022), "FinRL: a deep reinforcement learning library for automated stock trading in quantitative finance", *arXiv preprint*, available at: https://arxiv.org/abs/2011.09607

Markowitz, H. (1952), "Portfolio selection", *The Journal of Finance*, Vol. 7 No. 1, pp. 77-91, doi: 10.1111/j.1540-6261.1952.tb01525.x.

Mnih, V., Badia, A.P., Mirza, M., Graves, A., Lillicrap, T., Harley, T., Silver, D. and Kavukcuoglu, K. (2016), "Asynchronous methods for deep reinforcement learning", *Proceedings of the 33rd International Conference on Machine Learning (ICML)*, pp. 1928-1937.

Rezaei, M. and Nezamabadi-Pour, H. (2025), "A taxonomy of literature reviews and experimental study of deep reinforcement learning in portfolio management", *Artificial Intelligence Review*, Vol. 58 No. 3, p. 94, doi: 10.1007/s10462-024-11066-w.

Saberironaghi, M., Ren, J. and Saberironaghi, A. (2025), "Stock market prediction using machine learning and deep learning techniques: a review", *AppliedMath*, Vol. 5 No. 3, p. 76, doi: 10.3390/appliedmath5030076.

Schulman, J., Wolski, F., Dhariwal, P., Radford, A. and Klimov, O. (2017a), "Proximal policy optimization algorithms", *arXiv preprint*, available at: https://arxiv.org/abs/1707.06347

Schulman, J., Levine, S., Abbeel, P., Jordan, M. and Moritz, P. (2017b), "Trust region policy optimization", *Proceedings of the 32nd International Conference on Machine Learning (ICML)*, pp. 1889-1897.

Sutton, R.S. and Barto, A.G. (2018), *Reinforcement Learning: An Introduction*, 2nd ed., MIT Press.

Thi Bich, C.P., Ha, L.N. and Minh, T.V.T. (2025), "COVID-19 and liquidity dynamics: analyzing ASEAN-6 stock markets during and post-pandemic", *Sage Open*, Vol. 15 No. 2, pp. 1-18, doi: 10.1177/21582440251336112.

Velay, M., Doan, B.-L., Rimmel, A., Popineau, F. and Daniel, F. (2023), "Benchmarking robustness of deep reinforcement learning approaches to online portfolio management", *arXiv preprint*, available at: https://arxiv.org/abs/2306.10950

World Bank (2024), *Unlocking the Potential of Viet Nam's Capital Markets*, World Bank.

Corresponding author
Worrawat Saijai can be contacted at: worrawat.s@cmu.ac.th