Unstable Gains: Evaluating the Reproducibility of Deep Reinforcement Learning in Trading and Portfolio Management

Abstract:

This article investigates the reproducibility and robustness of deep reinforcement learning (DRL) in financial applications, focusing on algorithmic trading and portfolio management across two asset classes: stocks and cryptocurrencies. While DRL has gained popularity in these domains, most studies rely on single-run evaluations and overlook the high variance inherent to these methods. We reproduce influential DRL-based strategies under identical hyperparameters but across multiple independent random seeds and show that both performance and learned policies vary widely under fixed configurations. These experiments highlight the fragility of commonly reported results. Even the best performing algorithms display substantial variability across runs. To improve reliability, we introduce a checkpointing strategy and quantify uncertainty using bootstrapping and permutation tests. Our findings reveal that prevailing evaluation practices risk misleading conclusions about strategy efficacy and also conceal the true risk profile of DRL-based financial models. This underscores the need for more rigorous and reproducible protocols to ensure dependable advancements and foster genuine risk assessment in financial DRL research

Keywords

Deep Reinforcement Learning, Algorithmic Trading, Portfolio Management, Reproducibility, AI in Finance

1. Introduction

Artificial intelligence (AI) is transforming modern finance, with its profound impact increasingly evident in the investment sphere. A recent comprehensive review by Bahoo et al. (2024) underscores AI's expansive role, identifying ten major research streams where AI is applied. These streams cover diverse applications, with most directly pertaining to investment, such as stock market analysis, the development of trading models, portfolio management, the study of cryptocurrencies, and investor sentiment analysis. Further exemplifying AI's growing influence, sophisticated models are now being developed that can digest complex information from corporate disclosures to macroeconomic trends, in some cases reportedly surpassing human analysts in tasks such as stock return prediction (Cao et al., 2024). This growing body of work illustrates how AI is reshaping the way investment decisions are made, risks are assessed, and portfolios are optimized.

Among the diverse AI methodologies, Deep Reinforcement Learning (DRL) has emerged as a particularly compelling paradigm for financial applications such as algorithmic trading and portfolio management. Unlike traditional econometric or supervised machine learning methods, which typically require separate components or manual rules for tasks like signal generation, position sizing, and trade execution, DRL provides an integrated, end-to-end decision-making framework. A DRL agent learns directly from sequential market interactions, continuously updating its policy to make autonomous trading decisions in response to dynamic financial environments. This integration of signal discovery, action selection, and reward optimization within a single adaptive system makes DRL especially attractive for complex, data-rich financial applications. Notable studies—including Jiang et al. (2017), Liu et al. (2018) and Yang et al. (2020)—report promising results relative to benchmark strategies, leveraging actor-critic methods, ensemble techniques, and recurrent architectures. The literature has since

expanded rapidly, with hundreds of DRL-based models proposed for tasks such as portfolio management and order execution (Sun et al., 2023).

This rapid advancement underscores the enthusiasm for AI in finance. However, to ensure that this enthusiasm translates into reliable and trustworthy applications, it is crucial to critically evaluate the methodological underpinnings of these sophisticated tools. This paper focuses on DRL, arguing that despite its promise, prevailing research practices often overlook fundamental issues of reproducibility and robustness. Consequently, we challenge the claims of DRL's superiority in financial settings that often stem from these methodological shortcomings. It is well established in the reinforcement learning literature that DRL algorithms are highly sensitive to initial conditions, including random seeds, hyperparameters, and implementation details—even in relatively simple environments like MuJoCo. Foundational works by Henderson et al. (2017) and Islam et al. (2017), conducted in simulated non-financial environments, demonstrate that reported performance can vary substantially across runs, and that small implementation changes may lead to significantly different outcomes. Financial markets are even more complex: they are noisy, non-stationary, and difficult to simulate. Despite these challenges, most DRL studies in finance fail to quantify the variability of their results. This is true not only for some of the most widely cited papers, but also for recent work which claim superior performance of DRL (Huang et al., 2024; Y. Jiang et al., 2024; Li & Hai, 2024; Zou et al., 2024) or lack thereof (Kruthof & Müller, 2025). Notably, some studies (Jang & Seong, 2023; Majidi et al., 2024; Théate & Ernst, 2021) have begun to address this issue, while still claiming superiority of their innovations—an approach we believe is more responsible for reporting such results. Interestingly, the latter two studies report substantial dispersion in returns and Sharpe ratios within the same experiments, further validating our concerns. Recent surveys (Pricope, 2021; Sun et al., 2023) both describe financial-market DRL as a still-emerging area, noting that most studies stop at a single profit-focused back-test on ad-hoc data and baselines, with minimal attention to risk, robustness, or reproducibility. This practice hampers fair comparison with classical benchmarks and obscures the relative strengths of competing DRL algorithms, leaving the community without a clear consensus on the most effective methods for quantitative finance. A recent survey by Pippas et al. (2025) identifies the same weakness, documenting pervasive one-off back-tests, inflated Sharpe ratios and scant reporting of uncertainty and therefore reinforce our call for more robust results reporting.

In this paper, we replicate and extend prominent DRL trading strategies implemented in the FinRL library (Liu et al., 2021), which originally reported results from a single run per experiment. Our experiments cover two common financial tasks: algorithmic trading and portfolio management. As a key contribution, we repeat each experiment across twenty random seeds, revealing substantial variability in performance, even under otherwise identical settings. These findings highlight the methodological fragility of DRL in financial applications and underscore the need for more rigorous evaluation standards. We argue that future work should report results across multiple runs with confidence intervals and adopt standardized practices to improve robustness and reproducibility in financial DRL research.

2. Data and methodology

We use the open-source FinRL framework, which has contributed significantly to reproducibility in DRL research and has been adopted by other studies (Zou et al., 2024). Our analysis covers two distinct asset classes: U.S. equities and cryptocurrencies allowing us to represent both mature and emerging market segments and is inspired by influential studies we aim to reproduce. For the equity experiments, we use the 30 constituents of the Dow Jones Industrial Average, a diversified large-cap index representative of the U.S. market. The data spans January 2009 to December 2024, with a testing window from January 2021 onward,

capturing both bull and bear markets. For the cryptocurrency experiment, we select 11 widely traded digital assets with sufficient historical depth, all launched before 2018, and use daily price data from January 2018 to December 2024. All financial data are sourced from Yahoo Finance.

We design three experiments to evaluate DRL strategies in algorithmic trading and portfolio management. Each experiment is repeated 20 times using different random seeds, allowing us to assess the variability of results due to stochastic training processes. We quantify uncertainty using bootstrapping methods, following recommendations from Henderson et al. (2017).

In the first experiment, we replicate the influential Dow 30 ensemble trading strategy proposed by Yang et al. (2020). The second experiment compares various DRL algorithms and proposes a strategy to improve robustness within this framework, closely following the setup of the first experiment. The third experiment addresses portfolio management in the cryptocurrency market, based on the FinRL implementation of Jiang et al. (2017).

In the first two experiments, the state space is a 181-dimensional vector consisting of: available cash, closing prices of all stocks, the number of shares held, and four technical indicators—MACD, RSI, CCI, and ADX. The action space defines how many shares of each stock the agent can buy or sell, with an upper limit of 100 per timestep. The lower limit is set by the current holdings, implying no short selling. The reward is defined as the change in portfolio value between consecutive time steps. A transaction cost of 0.1% is applied to both buying and selling. Following Yang et al. (2020) we assume a risk-averse agent who moves entirely to cash when market volatility exceeds a predefined, data-driven threshold. We evaluate five widely used DRL algorithms:

- Deep Deterministic Policy Gradient (DDPG) (Lillicrap et al., 2015) is a deterministic, off-policy actor–critic algorithm for continuous action spaces that combines DQN-style target networks with an Ornstein–Uhlenbeck exploration process.
- Advantage Actor–Critic (A2C) (Mnih et al., 2016) is a synchronous version of A3C: multiple workers share parameters and compute the advantage-function baseline to lower policy-gradient variance.
- Proximal Policy Optimization (PPO) (Schulman et al., 2017) proposed a clipped surrogate objective for stable on-policy training with fewer tuning requirements.
- Twin Delayed Deep Deterministic Policy Gradient (TD3) (Fujimoto et al., 2018) Extends DDPG with twin critics, delayed policy updates, and target policy smoothing to mitigate Q-function overestimation.
- Soft Actor–Critic (SAC) (Haarnoja et al., 2018) is an off-policy actor–critic that maximizes reward plus an entropy term, producing stochastic policies that balance exploration and exploitation.

Experiment 2 shifts focus to the relative performance of individual DRL algorithms, refining the model selection approach within the rolling window framework. Experiment 1 also uses a rolling window, its selection mechanism involves choosing the single best-performing algorithm among the five candidates based on the validation period's results, deploying only that winner for the subsequent 63-day trading period. Recognizing that DRL training is volatile and final models may not be the best, Experiment 2 evaluates multiple saved checkpoints for each algorithm on the 63-day validation data before each trading period. The checkpoint yielding the highest Sharpe ratio for that specific algorithm is then selected and used for its trading during the next 63-day window. Furthermore, Experiment 2 employs hyperparameters optimized through an initial grid search, whereas Experiment 1 relied on default parameters from the library.

The third experiment implements a portfolio allocation environment (Costa & Costa, 2023). Instead of trading discrete shares, the agent assigns portfolio weights to cryptocurrencies and cash, constrained to sum to one, thus no short selling is allowed in this experiment either. The policy network is a convolutional architecture known as the Ensemble of Identical Independent Evaluators (EIIE), introduced by Jiang et al. (2017). It receives a 50×11×3 input matrix representing the last 50 timesteps, 11 assets, and three features: closing, high, and low prices. The EIIE design uses local receptive fields of size 1 in all feature maps, allowing the network to process each asset independently until the final softmax layer, which outputs allocation weights. To account for transaction costs (set to 0.25% as in the original paper), portfolio weights from the previous time step are appended in the later layers, allowing the agent to learn to minimize unnecessary reallocation. Reward function aims to maximize the average logarithmic cumulative return, the same as in the referenced study.

3. Empirical results

3.1. Experiment 1 - Stock Trading

In our first experiment, we replicate the ensemble-based DRL trading strategy proposed by Yang et al. (2020), using the 30 constituents of the Dow Jones Industrial Average as the investment universe. Our goal is to assess how training duration and stochastic variability inherent in DRL training affect reported performance, and to illustrate the risks of single-run evaluation that are common in financial DRL studies.

Figure 1 presents the average daily account value of the ensemble DRL strategy, along with its 95% bootstrapped confidence interval based on 20 independent training runs. For comparison, the performance of the Dow Jones Industrial Average is shown as a passive benchmark. The left panel shows results after 10,000 training steps, while the right panel uses 100,000 steps. We selected these durations to reflect the fact that many published studies do not specify stopping criteria or training duration—both of which can critically influence performance assessments. Experiments using 100,000 steps required several weeks to complete on an AWS ml.m5.xlarge instance, as our implementation uses multilayer perceptron (MLP) policies, which are not significantly accelerated by GPU hardware.

We find that a single DRL run may appear to outperform the index, particularly when trained for fewer steps. However, when considering the 95% bootstrap confidence interval across 20 runs, the index's trajectory often lies well within the model's uncertainty band. This suggests that any observed outperformance in single-run evaluations may be illusory, arising from stochastic elements such as random initialization or sampling noise during training. Relying on a single trajectory risks cherry-picking and does not reflect the true behavior or robustness of the strategy.

To assess statistical significance, we apply a non-parametric sign-flipping test to compare the average Sharpe ratio of the ensemble strategy to that of the buy-and-hold index. Under shorter training (10,000 steps), the difference is not statistically significant. Under longer training (100,000 steps), the DRL strategy performs significantly worse than the index ($p < 0.05$). These results highlight that longer training alone does not guarantee improved or reliable performance, and may in fact worsen it under some configurations.

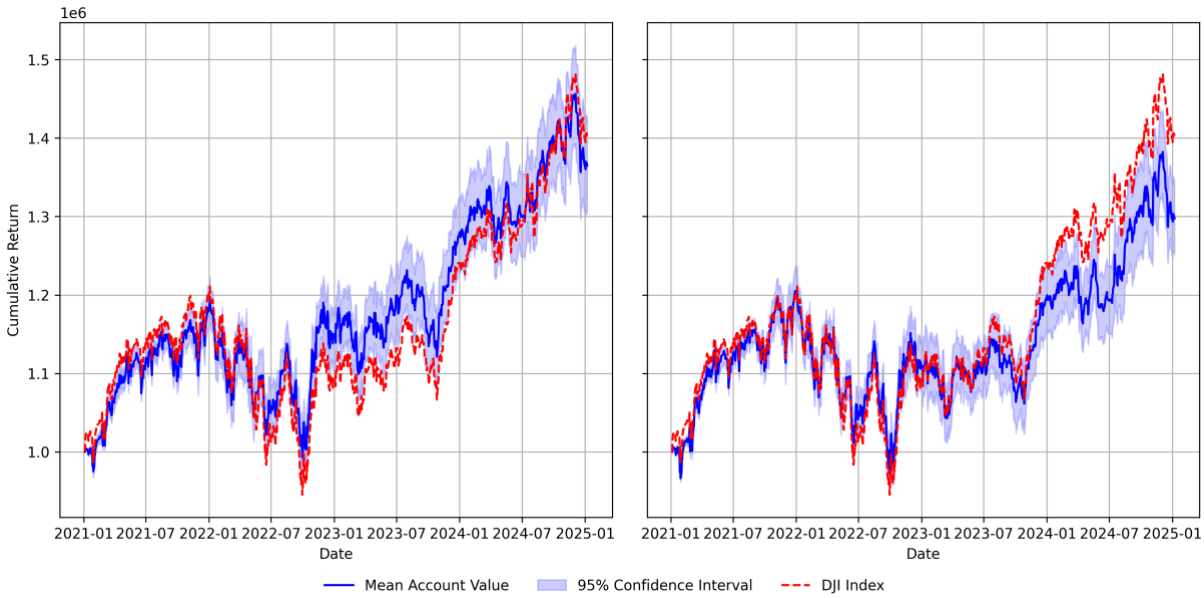

Figure 1. Ensemble strategy performance after 10,000 (left) and 100,000 (right) training steps per algorithm

Figure 2 further illustrates the instability of the learned policies. We report the average portfolio allocation across all 30 stocks, aggregated over the first ten simulations under 10,000 training steps, which produced comparatively better results. Despite using identical model configurations, the asset weights vary substantially across runs, suggesting that different policies are being learned each time. This variability violates the principle of algorithmic reproducibility, as proposed by Impagliazzo et al. (2023), which requires an algorithm to yield consistent outputs when trained on independent samples from the same distribution.

In financial applications, where stability and interpretability are essential, such policy-level volatility undermines the reliability of DRL-based trading strategies. If the same algorithm, trained under identical settings, recommends drastically different asset allocations depending on random initialization, then its use in real-world investment decisions becomes questionable. This experiment demonstrates that reporting a single successful run is not sufficient for evaluating DRL-based financial models, and may result in misleading conclusions about profitability or robustness.

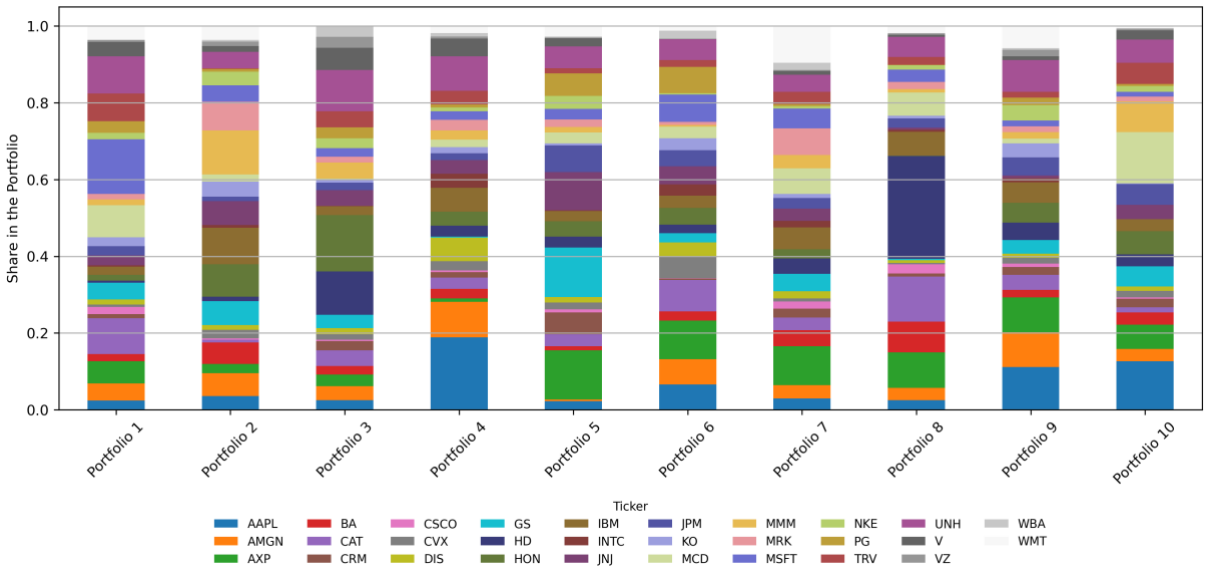

Figure 2. Average portfolio allocation across different seeds

3.2. Experiment 2 – Algorithms Comparison

A common limitation in financial DRL studies is the lack of clarity on how the final model is selected—specifically, the training duration, checkpointing strategy, and how many model configurations were tested. This lack of transparency can introduce selection bias, particularly when many runs are conducted but only the best results are reported (Bailey & Lopez de Prado, 2014). The issue is especially problematic in complex environments like trading, where reward trajectories during training are volatile and final models are not always optimal.

To address this issue, we conduct a second experiment that retains the 100,000 training steps used in Experiment 1 but introduces a checkpointing strategy. Each model is evaluated every 5,000 steps on a validation set, using the Sharpe ratio as a selection metric. The checkpoint achieving the highest Sharpe ratio is then used for trading in the subsequent period. This approach contrasts with the common practice of using the final model, which we found to be suboptimal in many cases. Additionally, in this experiment, we do not apply the turbulence threshold, allowing all policies to trade continuously and reflect full model behavior.

Figure 3 shows the average account value across 20 runs for each algorithm. As training progresses, performance dispersion between algorithms becomes more pronounced. Since each curve represents an average over 20 runs, we gain greater confidence in the relative rankings. Notably, DDPG consistently outperforms other methods, particularly A2C.

To assess whether these differences are statistically meaningful, we report 95% bootstrapped confidence intervals in Figure 4. While the intervals for DDPG and A2C overlap only marginally, formal testing is required to determine whether performance differences are statistically robust. We leveraged the pairing of seeds and applied a one-sided sign-flip permutation test to the per-seed Sharpe-ratio differences, followed by a Holm–Bonferroni correction for the four simultaneous contrasts. The raw p-values indicate that DDPG tends to outperform A2C ($p = 0.013$) and SAC ($p = 0.026$), but after family-wise error control the adjusted p-values rise to 0.052 and 0.078, respectively. Comparisons with PPO and TD3 are even less conclusive (adjusted $p > 0.10$). Consequently, no Sharpe-ratio difference remains statistically significant at the 5 % level. The largest effect size—a 0.16 Sharpe-unit edge over A2C—therefore cannot be deemed reliable. These findings illustrate how apparent wins can vanish once appropriate paired testing and multiple-comparison correction are applied. This focus on relative performance is central to many DRL studies, where new methods are benchmarked against established baselines. For example, Zhang et al. (2019) compare A2C and DQN across asset classes, reporting superior results relative to benchmarks. Yet, such claims should be interpreted cautiously when uncertainty is not quantified.

Interestingly, the best performing individual algorithms in this experiment outperform the ensemble strategy from Experiment 1. This suggests that selecting the best checkpointed model—rather than using the final model—can improve performance. However, we note that hyperparameter tuning may also have contributed to this effect.

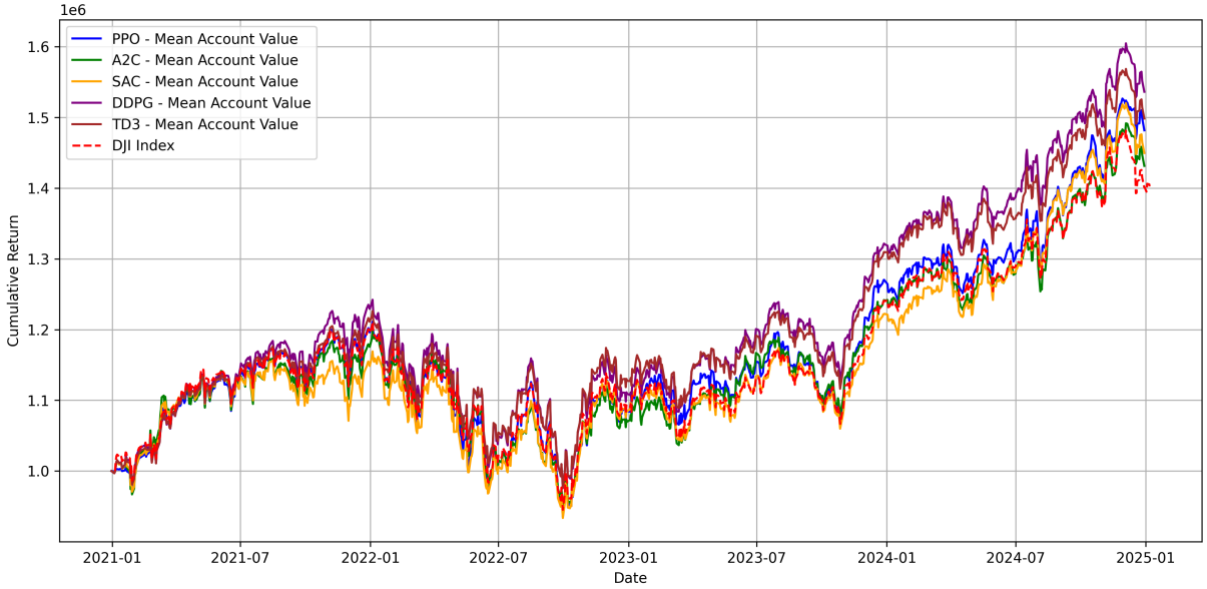

Figure 3. Average results from different DRL algorithms

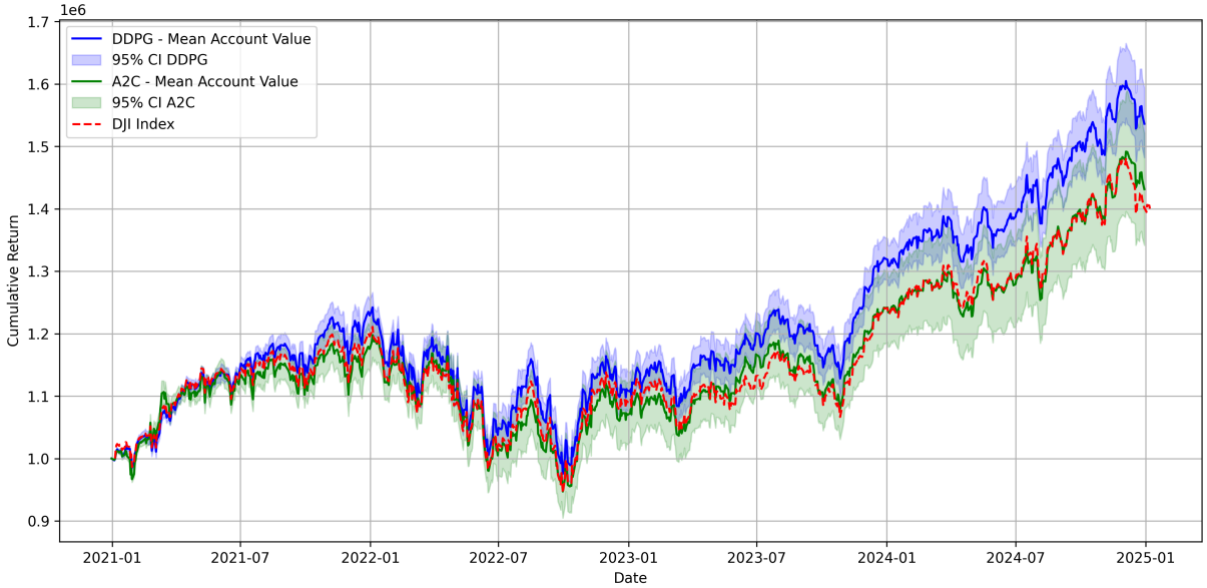

Figure 4. Confidence intervals of the best performing algorithm (DDPG) and the worst (A2C)

3.3. Experiment 3 – Cryptocurrencies portfolio management

In this experiment, we evaluate the performance of the agent in a cryptocurrency portfolio allocation setting. This market provides a natural stress test due to its high volatility**,** frequent regime shifts, and weaker market efficiency compared to traditional equities. We follow the setup of Jiang et al. (2017), using the EIIE architecture, and report results separately for each year from 2021 to 2024, in line with the format of the original study. The model was retrained independently each year for 50 episodes, as this setting yielded favorable Sharpe ratios on the training set and longer training durations produced similar results. We compare the agent's performance against two benchmarks: an equal-weighted portfolio of the eleven cryptocurrencies, and a fully Bitcoin-allocated portfolio**,** reflecting a more conservative strategy focused on the most widely held cryptocurrency.

Figure 5 summarizes annual performance across these four years. In 2021, a strong bull market, the DRL agent outperformed the Bitcoin benchmark but slightly underperformed the equal-weighted strategy. In the bear market of 2022, the agent showed more robust

performance, outperforming both benchmarks. During the market recovery in 2023 and 2024, the agent dynamically increased exposure to leading altcoins, outperforming the equal-weighted benchmark in 2023 and both benchmarks in 2024. These results suggest that the agent is capable of adapting to changing market conditions and reallocating capital toward outperforming assets.

However, consistent with our broader findings, these favorable outcomes must be interpreted with caution. The DRL agent's performance varied substantially across runs. For instance, in 2024, its best-performing year, the 95% bootstrapped confidence interval for annual returns ranged from approximately 200% to 400%, illustrating the high variance inherent in the training process. This reinforces our position that conclusions drawn from a single run are unreliable. Reported gains may reflect favorable randomness rather than consistent algorithmic advantage. Thus, in financial applications, particularly in volatile markets like cryptocurrencies, comparison against robust baselines and uncertainty quantification are essential prerequisites for making credible performance claims.

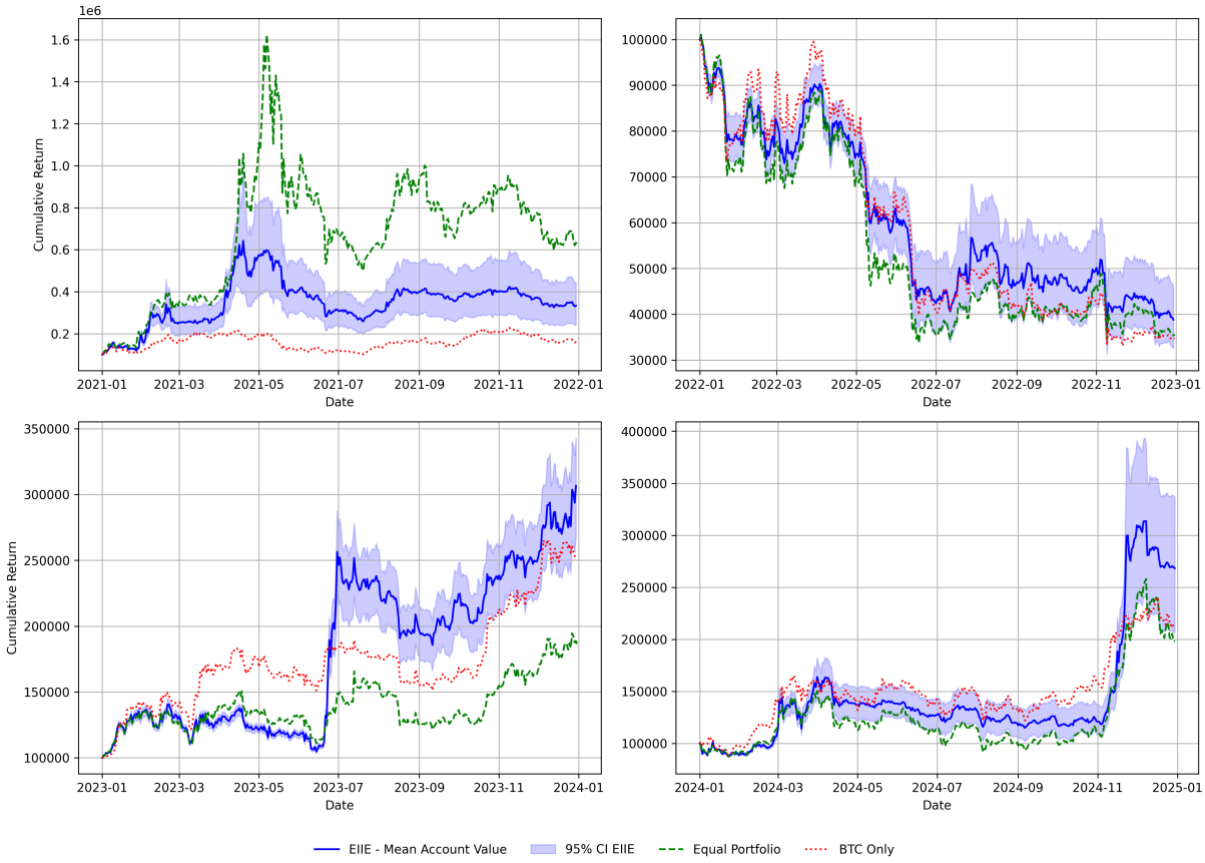

Figure 5. EIIE results for cryptocurrency portfolio management

4. Conclusions

This study exposes the methodological fragility of deep reinforcement learning (DRL) in financial applications. Across three experiments—stock trading with ensemble strategies, comparative evaluation of DRL algorithms, and cryptocurrency portfolio management—we demonstrate that performance varies substantially across random seeds, training durations, and model selection strategies, even under fixed hyperparameters.

Our findings underscore a core limitation of current DRL practices in finance: results based on single runs are unreliable and often overstate algorithmic performance. Moreover, they conceal the true risk profile and instability inherent in these complex models. We show that

incorporating bootstrapped confidence intervals, permutation-based significance tests, and multiple independent runs provides a more honest and statistically grounded assessment. Even when average returns are high, they are frequently accompanied by wide confidence intervals and unstable policy behavior, limiting their practical reliability.

To improve evaluation reliability, we also introduce checkpointing, selecting models based on their validation-set Sharpe ratios rather than defaulting to final weights. This approach, though standard in supervised learning, remains underused in DRL and leads to improved out-of-sample performance.

We advocate for a shift in evaluation standards. Future research, particularly involving DRL but also extending its principles to other complex machine learning models used in finance**,** should adopt standardized, transparent protocols that prioritize robustness and reproducibility over potentially misleading point estimates. This includes mandatory reporting of results over multiple seeds, comprehensive uncertainty quantification, and clear documentation of all model selection procedures. As financial applications increasingly rely on sophisticated AI-driven models, ensuring their reproducibility, stability, and statistical integrity is not merely a matter of scientific credibility, but a prerequisite for responsibly unlocking AI's full potential in reshaping finance and for building trust in real-world investment and risk management contexts.

Declaration of generative AI and AI-assisted technologies in the writing process
During the preparation of this work the authors used ChatGPT in order to improve language and readability. After using this tool/service, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

Data availability

Data is publicly available.

Bibliography

Bahoo, S., Cucculelli, M., Goga, X., & Mondolo, J. (2024). Artificial intelligence in Finance: A comprehensive review through bibliometric and content analysis. *SN Business & Economics*, *4*(2), 23. https://doi.org/10.1007/s43546-023-00618-x

Bailey, D. H., & Lopez de Prado, M. (2014). *The Deflated Sharpe Ratio: Correcting for Selection Bias, Backtest Overfitting and Non-Normality* (SSRN Scholarly Paper No. 2460551). Social Science Research Network. https://doi.org/10.2139/ssrn.2460551

Cao, S., Jiang, W., Wang, J., & Yang, B. (2024). From Man vs. Machine to Man + Machine: The art and AI of stock analyses. *Journal of Financial Economics*, *160*, 103910. https://doi.org/10.1016/j.jfineco.2024.103910

Costa, C. de S. B., & Costa, A. H. R. (2023). POE: A General Portfolio Optimization Environment for FinRL. *Brazilian Workshop on Artificial Intelligence in Finance (BWAIF)*, 132–143. https://doi.org/10.5753/bwaif.2023.231144

Fujimoto, S., Hoof, H. van, & Meger, D. (2018). *Addressing Function Approximation Error in Actor-Critic Methods* (No. arXiv:1802.09477). arXiv. https://doi.org/10.48550/arXiv.1802.09477

Haarnoja, T., Zhou, A., Abbeel, P., & Levine, S. (2018). *Soft Actor-Critic: Off-Policy Maximum Entropy Deep Reinforcement Learning with a Stochastic Actor* (No. arXiv:1801.01290). arXiv. https://doi.org/10.48550/arXiv.1801.01290

Henderson, P., Islam, R., Bachman, P., Pineau, J., Precup, D., & Meger, D. (2017). *Deep Reinforcement Learning that Matters* (No. arXiv:1709.06560). arXiv. https://doi.org/10.48550/arXiv.1709.06560

Huang, Y., Wan, X., Zhang, L., & Lu, X. (2024). A novel deep reinforcement learning framework with BiLSTM-Attention networks for algorithmic trading. *Expert Systems with Applications*, *240*, 122581. https://doi.org/10.1016/j.eswa.2023.122581

Impagliazzo, R., Lei, R., Pitassi, T., & Sorrell, J. (2023). *Reproducibility in Learning* (No. arXiv:2201.08430). arXiv. https://doi.org/10.48550/arXiv.2201.08430

Islam, R., Henderson, P., Gomrokchi, M., & Precup, D. (2017). *Reproducibility of Benchmarked Deep Reinforcement Learning Tasks for Continuous Control* (No. arXiv:1708.04133). arXiv. https://doi.org/10.48550/arXiv.1708.04133

Jang, J., & Seong, N. (2023). Deep reinforcement learning for stock portfolio optimization by connecting with modern portfolio theory. *Expert Systems with Applications*, *218*, 119556. https://doi.org/10.1016/j.eswa.2023.119556

Jiang, Y., Olmo, J., & Atwi, M. (2024). Deep reinforcement learning for portfolio selection. *Global Finance Journal*, *62*, 101016. https://doi.org/10.1016/j.gfj.2024.101016

Jiang, Z., Xu, D., & Liang, J. (2017). *A Deep Reinforcement Learning Framework for the Financial Portfolio Management Problem* (No. arXiv:1706.10059). arXiv. https://doi.org/10.48550/arXiv.1706.10059

Kruthof, G., & Müller, S. (2025). Can deep reinforcement learning beat 1/N. *Finance Research Letters*, *75*, 106866. https://doi.org/10.1016/j.frl.2025.106866

Li, H., & Hai, M. (2024). Deep Reinforcement Learning Model for Stock Portfolio Management Based on Data Fusion. *Neural Processing Letters*, *56*(2), 108. https://doi.org/10.1007/s11063-024-11582-4

Lillicrap, T. P., Hunt, J. J., Pritzel, A., Heess, N., Erez, T., Tassa, Y., Silver, D., & Wierstra, D. (2015). *Continuous control with deep reinforcement learning* (No. arXiv:1509.02971). arXiv. https://doi.org/10.48550/arXiv.1509.02971

Liu, X.-Y., Yang, H., Gao, J., & Wang, C. (2021). *FinRL: Deep Reinforcement Learning Framework to Automate Trading in Quantitative Finance* (SSRN Scholarly Paper No. 3955949). https://doi.org/10.2139/ssrn.3955949

Majidi, N., Shamsi, M., & Marvasti, F. (2024). Algorithmic trading using continuous action space deep reinforcement learning. *Expert Systems with Applications*, *235*, 121245. https://doi.org/10.1016/j.eswa.2023.121245

Mnih, V., Badia, A. P., Mirza, M., Graves, A., Lillicrap, T. P., Harley, T., Silver, D., & Kavukcuoglu, K. (2016). *Asynchronous Methods for Deep Reinforcement Learning* (No. arXiv:1602.01783). arXiv. https://doi.org/10.48550/arXiv.1602.01783

Pippas, N., Turkay, C., & Ludvig, E. A. (2025). *The Evolution of Reinforcement Learning in Quantitative Finance: A Survey* (No. arXiv:2408.10932). arXiv. https://doi.org/10.48550/arXiv.2408.10932

Pricope, T.-V. (2021). *Deep Reinforcement Learning in Quantitative Algorithmic Trading: A Review* (No. arXiv:2106.00123). arXiv. https://doi.org/10.48550/arXiv.2106.00123

Schulman, J., Wolski, F., Dhariwal, P., Radford, A., & Klimov, O. (2017). *Proximal Policy Optimization Algorithms* (No. arXiv:1707.06347). arXiv. https://doi.org/10.48550/arXiv.1707.06347

Sun, S., Wang, R., & An, B. (2023). Reinforcement Learning for Quantitative Trading. *ACM Transactions on Intelligent Systems and Technology*, *14*(3), 1–29. https://doi.org/10.1145/3582560

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

Yang, H., Liu, X.-Y., Zhong, S., & Walid, A. (2020). *Deep Reinforcement Learning for Automated Stock Trading: An Ensemble Strategy* (SSRN Scholarly Paper No. 3690996). https://doi.org/10.2139/ssrn.3690996

Zhang, Z., Zohren, S., & Roberts, S. (2019). *Deep Reinforcement Learning for Trading* (No. arXiv:1911.10107). arXiv. https://doi.org/10.48550/arXiv.1911.10107

Zou, J., Lou, J., Wang, B., & Liu, S. (2024). A novel Deep Reinforcement Learning based automated stock trading system using cascaded LSTM networks. *Expert Systems with Applications*, *242*, 122801. https://doi.org/10.1016/j.eswa.2023.122801