acrt AI, Computer Science and Robotics Technology

IntechOpen Journals

RESEARCH PAPER

Deep Reinforcement Learning Approach to Portfolio Optimization in the Australian Stock Market

Weiye Wu and Carol Anne Hargreaves*

Department of Statistics and Data Science, Faculty of Science, National University of Singapore, Singapore
*Corresponding author. E-mail: carol.hargreaves@nus.edu.sg

Citation
Weiye Wu and Carol Anne Hargreaves (2024), Deep Reinforcement Learning Approach to Portfolio Optimization in the Australian Stock Market. *AI, Computer Science and Robotics Technology* 3(1), 1–31.

DOI
https://doi.org/10.5772/acrt.20230095

Received: 27 October 2023
Accepted: 28 August 2024
Published: 19 September 2024

Abstract

The future of portfolio management is evolving from relying on human expertise to incorporating artificial intelligence techniques. Traditional techniques such as fundamental and technical analysis will eventually be replaced by more sophisticated deep reinforcement learning (DRL) algorithms. However, it is still a long way from designing a profitable strategy in the complex and dynamic stock market. While previous studies have focused on the American stock market, this paper applies two DRL algorithms, the proximal policy optimization (PPO) and the advantage actor–critic (A2C), to trade the constituent stocks of the Australian Securities Exchange 50 (ASX50) Index. This paper also incorporates a weighted moving average into the action space and introduces a transaction threshold to help agents minimize trivial trades that lead to high transaction costs. The results are presented and benchmarked against the ASX50 Index. The A2C agent was better at following trends and had the higher upside potential but can suffer from more severe damage during bearish markets. On the other hand, the PPO agent had the lowest annual volatility and the highest maximum drawdown, which is more helpful in a bearish or volatile market.

Keywords: deep reinforcement learning, portfolio optimization, proximal policy optimization, advantage actor–critic

1. Introduction

Portfolio management is a time-honored topic that has been widely discussed and studied in the modern financial world. Traditionally, portfolio management has relied on expert human knowledge on financial markets, asset valuation, and risk management, which was primarily the domain of wealthy individuals and institutions.

However, the advent of online brokerages and social media greatly lowered the barrier of entry into the world of investment. Online brokerages are offering low commission cost and benefits to new investors. Additionally, sharing of trading strategies and successes on social media has brought a lot of attention to the stock market. On top of that, the perfect storm of pandemic-induced boredom and disposable income from reduced spending further drives the average person's interest to an all-time high. Brokerages are reporting record new retail investors and trading volume [1], with assets worth billions of dollars being traded daily.

Even though the stock market is more accessible to the average person now, it still remains complex and deciding on what stocks to invest still remain an intimidating question. The common techniques to help investors with their decisions include fundamental analysis, which measures the intrinsic value of a stock using macro- and microeconomic factors; technical analysis, which evaluates a stock using statistical trends related to price and volume; and algorithmic trading [2] with machine learning techniques.

With remarkable progress in the artificial intelligence field, it is no surprise that about 75% of the trading volume on the United States stock market is made through algorithmic trading [2]. The main challenges to algorithmic trading are extracting information from noisy stock data and designing a suitable trading strategy [3]. Many studies have attempted to forecast the stock market by leveraging the use of sophisticated deep learning methods for their strong feature extraction ability [4–6]. However, such models lack decision-making capabilities [7]. These models focus on stock price forecasting. The trading strategies are then developed based on pre-defined rules and instructions. Stock prices are influenced by many unpredictable factors, from economic policies to natural disasters etc. [8], which will lower the effectiveness of such static trading strategies. Further, Zulfiquar *et al.* [9] performed research on the validity of the efficient market hypothesis and reaffirmed the existence and profitability of momentum investment strategies in 40 countries around the world during the period 1996–2018. In [9], the findings were robust to two distinct subperiod analyses, and there was a clear rejection of the efficient market hypotheses. This research outcome was valuable to momentum traders and stock market regulators.

Most classical reinforcement learning (RL) algorithms do not consider the exogenous nature and noise of financial time series data, which may lead to treacherous trading decisions. To address this issue, Yue *et al.* [10] proposed a novel anti-risk portfolio trading method based on deep reinforcement learning (DRL). It consists of a stacked sparse denoising autoencoder network and an actor–critic-based RL agent. Reinforcement learning is an autonomous self-teaching system that aims to develop an agent capable of making decisions through trial and error. The stock market can be modeled as a sequential decision-making process that can be navigated using an RL approach. However, the sudden and rapidly changing nature of the stock market makes it challenging for the agent to adapt and will eventually overwhelm the agent.

Therefore, deep reinforcement learning, a combination of deep learning and reinforcement learning, has emerged as a promising approach to portfolio management. The adoption of deep learning techniques helps the RL algorithms scale and manage big and noisy data. In recent years, there has been a growth in the studies applying DRL to portfolio management [10–12]. Nevertheless, these studies are not without some limitations.

First, the American stock market is heavily studied, especially the constituent stocks in the Dow Jones Industrial Average. This has limited the scope of the studies. Second, historical stock data between 2000 and 2020 was commonly used to train and test the model. The stock market has undergone major transformations over the past decade and is expected to continue evolving. The DRL models have a strong tendency to overfit [12]; thus, using data that are too old may be detrimental to the model.

It will be more interesting to study the application of DRL on the non-American stock market within the past 5 years. Therefore, this paper studied the daily stock data of the constituent stocks in the Australian Securities Exchange 50 (ASX50) Index from 2017 to 2022 and applied two DRL algorithms, namely advantage actor–critic (A2C) and proximal policy optimization (PPO), to manage the portfolio. This paper introduces limitations on the agent's action space through a weighted moving average and a transaction threshold to mitigate risks and enforce specific strategies that align with prior knowledge. However, this approach restricts autonomy of the agent and its ability to explore novel strategy organically. Thus, this paper explores how such interventions would influence the adaptability of the DRL algorithms.

The remainder of this paper is organized as follows. Section 2 reviews the related works of this study. Section 3 defines the portfolio management problem. Section 4 describes the methodology of the study and the necessary backgrounds. Section 5 presents the results of the models, which are validated through analysis. Finally, Section 6 concludes the paper and discusses interesting leads as future works.

2. Related work

Studies related to stock trading often involves large sums of money. Private firms are secretive about their study results [12]. As such, state-of-the-art models and promising results are not publicly available. Nevertheless, there is still study accessible on the topic of portfolio management. This section discusses the conventional approach to portfolio management and the recent works in the DRL approach to the problem.

2.1. Conventional methods

The volatility and uncertainty in the stock market will introduce risk to any investment in a stock portfolio. There is no single solution to portfolio management. Modern portfolio theory or mean-variance analysis [13], is one of the prominent methods used for portfolio management. The model assumes investors are risk-averse and rational, aiming to diversify the portfolio by minimizing the variance given a certain level of expected return [14]. The portfolio with the minimal variance is identified: this is called the minimum-variance portfolio. However, the minimum-variance portfolio does not account for market fluctuations and sees the market from a very long-term view. Future asset volatility is difficult to ascertain in practice [15]. This results in a considerable amount of estimation error and unstable composition of the portfolio.

Another well-known method for portfolio management is momentum trading. Investors identify the inertia of a price trend to seek profit from the herding behavior of market psychology. Momentum effects can be observed in stock markets across the world and are not limited to any particular market [14]. However, momentum strategies suffer heavily from sudden changes in market volatility. Any changes in price trends will inflict significant losses for investors trading based on the momentum.

Overall, conventional methods rely on static assumptions about the market such as the market is efficient or asset returns and risks can be accurately estimated [16]. These assumptions do not accurately reflect the dynamic nature of the market. Thus, the conventional models lack adaptability and is profitable only in the very long run.

2.2. Deep reinforcement learning methods

The DRL algorithms, on the other hand, are able to learn directly from raw sensory input without much domain-specific knowledge. The DRL aims to approximate optimal value or policy functions using deep neural networks and feedback signals from the environment. For example, Brini *et al.* [17] considered DRL agents that leverage classical strategies to increase their performances, and showed that this

approach guarantees flexibility, outperforming the benchmark strategy when the price dynamics was misspecified and some original assumptions on the market environment were violated in the presence of extreme events and volatility clustering. In addition, Chaoki *et al.* [18] demonstrated the potential of model-free RL methods to derive trading policies. The environments were chosen such that an optimal or close-to-optimal trading strategy was known. Their research showed that an RL agent can successfully recover the essential features of the optimal trading strategies and achieve close-to-optimal rewards. The DRL approaches can be divided into three main categories: critic-only, actor-only, and actor–critic.

2.2.1. *Critic-only approach*

The core idea of a critic-only or value-based method is to estimate the state–action values based on the Q-value function using a neural network. An investment decision will then be made by selecting the action with the highest predicted Q-value.

Researchers such as Chen *et al.*, Dang, and Li *et al.* [3, 7, 19–21] applied variations of the Deep Q-Network (DQN), which is the classic critic-only model, and were able to outperform conventional methods. Brim *et al.* [22] generated candlestick images as input to the convolutional neural network within a double DQN model. Li *et al.* [21] explored the inclusion of market sentiment as an input to the DQN model.

The limitation of the critic-only approach stems from its inherent reliance on discrete state and action space, which limits the actions that the agent can take. Adapting DQN to work with continuous action spaces in the case of portfolio management is challenging due to its reliance on discrete Q-values for each possible action.

2.2.2. *Actor-only approach*

On the other hand, an actor-only or policy-based method is where the agent learns the optimal policy directly without the need to compute and compare expected outcomes of actions by generating an optimal probability distribution over the possible actions. An investment decision will then be made by selecting the action with the highest probability.

Wang *et al.* [11] proposed a portfolio generator with a market scoring unit that embeds the macro market conditions and an asset scoring unit that evaluates the individual stock. The policy gradient algorithm is then applied to optimize the trading policy. In addition, Liang *et al.* [23] presented the performances of the policy gradient under different settings, including different learning rates, objective functions, and feature combinations in order to provide insights into parameter tuning, feature selection, and data preparation.

The actor-only model is not a popular approach in portfolio management. The limitation of the actor-only approach is the tendency to converge to a local optimum, which results in high variance when evaluating the policy. Although there are techniques to reduce this variance, there is a better approach to portfolio management: the actor–critic approach.

2.2.3. Actor–critic approach

The actor–critic framework is made up of two models: the actor network representing the policy that decides the action to take in a given state and the critic network representing the value function that evaluates the action that was taken. The actor–critic approach is able to combine the strengths of the critic-only and the actor-only model, making it the most popular approach for portfolio management.

Yang *et al.* [24] proposed an ensemble trading strategy using three actor–critic algorithms: PPO, A2C, and DDPG. The three models are validated using a 3-month rolling window to select the best-performing agent with the highest Sharpe ratio (SR). The selected agent would trade for the next interval. A turbulence index was included in the model to detect any extreme market conditions. The agents were able to sell off all the currently held stocks in the 2020 stock market crash and waited for the market to return to normal before trading again.

Sadriu [25] applied A2C and DDPG to optimize a stock portfolio trading 28 constituent stocks of the OMX Stockholm Index. The paper evaluated the performance of the two models during the 2020 stock market crash and noticed that the conventional methods actually outperformed the DRL models. Although the DRL models could recover from the crash faster, this shows the inability of the DRL models to respond to extreme conditions in the stock market.

Instead of using the usual open, high, low, close price, and volume (OHLCV) and technical indicators of the stocks, Yue *et al.* [10] proposed an anti-risk portfolio trading method based on the A2C algorithm with a sparse denoising autoencoder to first reconstruct the original financial data to minimize the impact of external factors such as political news, natural disasters, etc. Koratamaddi *et al.* [26] proposed a sentiment-award approach as an extension to the DDPG. The market sentiment score is first calculated using headlines from Google News and Twitter and passed as input to the model. Zhang *et al.* [27] experimented with both a discrete action space using a DQN and the policy gradient algorithm and a continuous action space with an A2C algorithm.

3. Problem definition

Portfolio management is the process of continuous reallocation of capital to create and maintain a diversified portfolio. We modeled portfolio management process as a

Markovian decision process (MDP). An MDP is defined as the tuple (S, A, p, r) consisting of the following:

- S as the state space, which includes the stock price information and the composition of the current portfolio
- A as the action space, which is the desired composition of the portfolio
- $p(s_t, a_t, s_{t+1})$ as the probability of transitioning to s_{t+1} from state s_t by taking action a_t
- $r(s_t, a_t, s_{t+1})$ as the immediate reward obtained from taking action a_t at state s_t and transitioning to state s_{t+1}.

The goal of the trading agent was to find an optimal policy π^* that determines the best course of action to take at any given state. There are three assumptions made about the portfolio management process:

- The liquidity of the market is high enough that there will be zero slippage. This implies that each trade the trading agent makes will be carried out at the market price.
- The capital invested by the trading agent is not significant enough to have any influence on the market.
- All transactions will be made at the closing price.
- Short selling is not allowed.

4. Materials and methods

4.1. Dataset description

The ASX50 Index consists of the top 50 companies listed on the Australian Stock Exchange based on market capitalization [28]. The index is weight-adjusted by the market capital of the constituent stocks. These stocks have a high trading volume, which suggests higher market liquidity, justifying our first assumption.

We extracted the daily OHLCV data of the stocks between 2017/01/01 and 2023/01/06 from EODData. The stocks selected were from the index composition at the end of the time period. Although there have been changes to the constituent stocks in the index over the 5-year period, it is not in our interest to mimic the index.

Although there are 50 companies in the ASX50 Index, 5 companies had been listed on the exchange later than 2017 and did not have data for the whole period. We removed these stocks from the portfolio. Therefore, the final dataset consisted of 45 stocks in total (see Appendix A).

Figure 1 shows the closing price of the ASX50 Index over the past 5 years. The ASX50 Index has fluctuated wildly over the past 5 years: strong growth in 2019 followed by the COVID-19 crash in the first half of 2020. The index has recovered steadily since the crash but is experiencing some volatility in 2022.

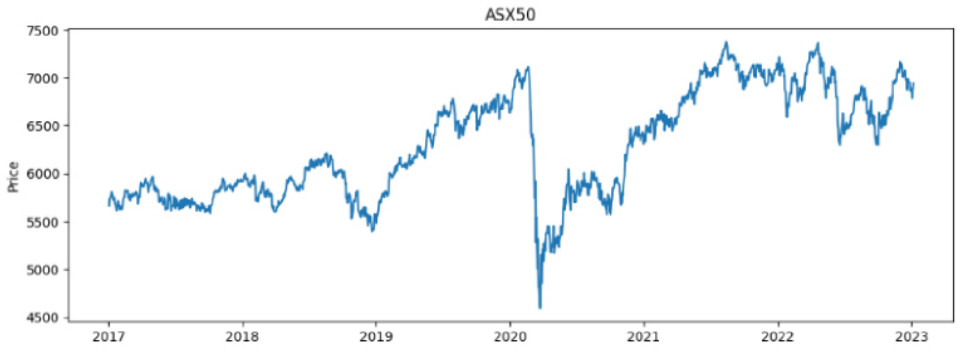

Figure 1. Closing price of the ASX50 Index (2017/01/01–2023/01/06).

The dataset was split into three periods as shown in Figure 2. Data from 01/01/2017 to 12/31/2019 were used for training, while data from 01/01/2020 to 12/31/2020 were used for validation. Finally, we would test our agent's performance on data from 01/01/2021 to 06/01/2023.

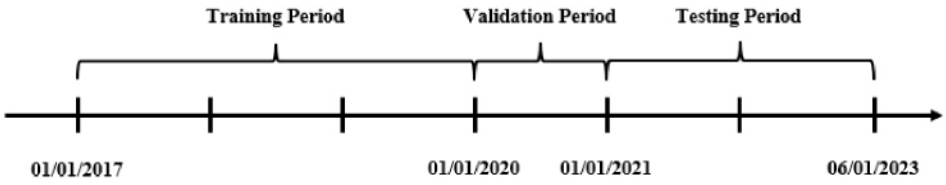

Figure 2. Training, validation, and testing periods.

4.1.1. Technical indicators

Apart from the OHLCV data, we computed the financial technical indicators to help describe the market. Technical indicators reflect the stock market movements from a different perspective and reduce the impact of noise on the data [29]. The technical indicators fit into four categories: trend, momentum, volume, and volatility. Table 1 shows a summary of the technical indicators used. (See Appendix B for the details.)

4.2. Portfolio management environment

Our portfolio management environment was built using OpenAI Gym [30], an open-source Python library for developing RL algorithms. It is important to build an

Table 1. Summary of financial technical indicators.

Category	Financial technical indicators
Trend	5-day exponential moving average (EMA)
Trend	20-day exponential moving average (EMA)
Momentum	Moving average convergence divergence (MACD)
Momentum	Relative strength index (RSI)
Momentum	Stochastic oscillator
Volume	On-balance volume (OBV)
Volume	Volume price trend (VPT)
Volume	Money flow index (MFI)
Volatility	Bollinger Bands

environment that simulates real-world trading as closely as possible so that the optimal policy learned can be applicable in the real world.

4.2.1. State space

The state space of our portfolio management environment at any given time t consisted of two components: a 3D tensor (P_t) with the OHLCV and technical indicator information and a vector (w_t) that represents the current portfolio weight. The tensor P_t has dimensions (n, w, f), where $n = 45$ is the number of available stocks, $w = 10$ is the length of the observations in days, and $f = 17$ is the number of features (OHLCV and technical indicators). A 10-day window was chosen based on the idea that recent market trends and momentum changes hold greater significance for decision making [31]. The observations were then normalized using MinMaxScaler from scikit-learn. The transformation is given by

$$x_{\text{scaled}} = \frac{x - \min(x)}{\max(x) - \min(x)}. \tag{1}$$

Here, $w_t = (w_{t,0}, w_{t,1}, \ldots, w_{t,n})^{\mathrm{T}}$ has length $n + 1$, where $w_{t,0}$ is the proportion of cash and $\forall i > 0$, $w_{t,i}$ is the proportion of stock i in the portfolio at time t. The elements of w_t will always sum to 1 by definition. Then w_t is initialized to $(1, 0, 0, \ldots, 0)^{\mathrm{T}}$, indicating that all the capital is in cash.

4.2.2. Action space

The action space of our agent was designed to handle continuous allocation of capital across a portfolio of assets, including cash:

1. Action vector: $a_t = (a_{t,0}, a_{t,1}, \ldots, a_{t,n})$ has length $n + 1$, where $a_{t,0}$ is the proportion of cash and $\forall i > 0$, $a_{t,i}$ is the proportion of capital allocated to stock i.
2. The elements of a_t must sum to 1: $\sum_{i=0}^{n} a_{t,i} = 1$.
3. No short selling of the stock: $0 \le a_{t,i} \le 1$, $\forall i$.

The policy output would consist of a vector of length $n + 1$ that directly represents the allocation strategy of the agent based on its observations of the environment.

4.2.3. Initial balance and transaction cost

In order to minimize the impact of the trades on the stock market, we gave the agent a smaller initial balance of $50,000. This is also a more reasonable amount for retail investors to start with.

Different brokerage platforms charge different amounts of commission fees. Given the small starting capital the agent has, the agent would be incurring the minimum fee for each trade. Hence, a fixed commission fee of $20 per trade was introduced, instead of a linear or quadratic cost model.

The introduction of a transaction fee should deter greedy algorithms that aim at achieving an optimal portfolio at every time t [23]. The transaction fee is very much present in the real world and can be very impactful to retail investors with low capital.

4.2.4. Reward function

The ultimate goal of all investors would be to maximize their returns. Therefore, we set the reward function as the logarithmic rate of returns:

$$r_t = \ln\left(\frac{p_t}{p_{t-1}}\right) \tag{2}$$

where p_t is the net returns at time t.

Saturation and inefficiency in learning can be an issue if the output is too sparse. When the reward function has a large range of values, certain actions may be assigned unnecessarily high importance, which can be disadvantageous in the long run. As such, we applied a reward scale of 0.1. This will compress the space of estimated expected returns and prevent any numerical instability [8].

4.3. Deep reinforcement learning models

We applied PPO and A2C algorithms from the Stable-Baselines3 library, which is a set of pre-built RL algorithms with the TensorFlow library [32].

4.3.1. Proximal policy optimization

The PPO is an on-policy algorithm that optimizes the policy function while ensuring that the distance between the new and old policies is not too large. This constraint helps to prevent the policy from changing too much in each iteration, which could lead to instability or suboptimal convergence.

The PPO uses a surrogate objective function to approximate the objective function. The surrogate objective function restricts the range within which the

policy can change. The clipped ratio is introduced to clip the policy update to a range between $1 - \varepsilon$ and $1 + \varepsilon$, where ε is a small positive number.

4.3.2. Advantage actor–critic

The A2C is an on-policy algorithm and an extension to the baseline actor–critic algorithm. In the baseline actor–critic algorithm, there are two models: an actor network that generates the policy and a critic network that measures how good the chosen action is. However, the log probability in the objective function has very high variance as it involves taking logarithms of small probabilities.

Instead of the action value function, A2C uses an advantage function. The idea of the advantage function is to determine how much better a certain action is compared to the average action. The advantage function provides a baseline that leaves behind only the part that is attributable to the action. This helps to mitigate the high variance in the policy gradients and ensure more stable training.

4.4. Performance metrics

We looked at five metrics to evaluate the performance of the agents. This includes compound annual growth rate (CAGR), cumulative return, annual volatility, Sharpe ratio, and maximum drawdown (MDD).

The CAGR is the measure of the annual return of an investment over time. The CAGR can be expressed as follows:

$$\text{CAGR} = \left(\frac{V_{\text{final}}}{V_{\text{initial}}}\right)^{\frac{1}{t}} - 1 \tag{3}$$

where V_{final} is the final value, V_{initial} is the initial value, and t is time in years.

Cumulative return is the cumulative sum of the daily returns and can be expressed as follows:

$$\text{CR} = \frac{V_{\text{final}} - V_{\text{initial}}}{V_{\text{initial}}}. \tag{4}$$

Annual volatility represents the dispersion of the returns and can be expressed as follows:

$$\sigma_p = \sqrt{252} \times \sigma \tag{5}$$

where σ is the standard deviation of the returns.

The SR is the measure of risk-adjusted return, which describes how much excess return you are receiving for the volatility incurred. We defined risk-free rate to be the Australia 10-Year Government Bond in this study. The SR can be expressed

as follows:

$$\text{SR} = \frac{R_p - R_f}{\sigma_p} \tag{6}$$

where R_p is the return of portfolio and R_f is the risk-free rate.

The MDD is the maximum drop in the value of the investment by finding the value between the peak and a trough before a new peak is attained.

5. Results

As shown in Figure 3, the performance of both the PPO and A2C fell short of expectations with the returns constantly plunging over time and eventually hitting rock bottom. Further investigations suggest that the underachieving results were due to the high volume of daily transactions in the first quarter of 2020.

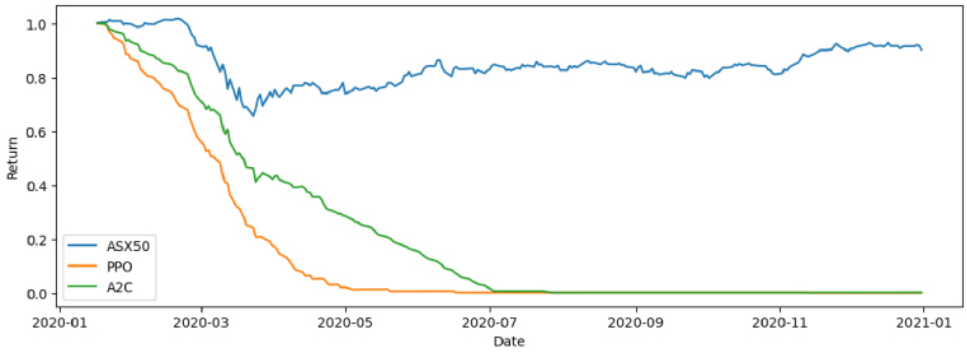

Figure 3. Cumulative return for PPO and A2C portfolios against ASX50. Both agents were unable to overcome the steep transaction cost.

Figures 4 and 5 show the distribution and trend of number of transactions made per day by the PPO agent, respectively. The PPO agent made around 30.67 transactions, paying a transaction cost of $613.40 per day. Figures 6 and 7 show the distribution and trend of number of transactions made per day by the A2C agent, respectively. The A2C agent made around 18.27 transactions, paying a transaction cost of $365.40 per day.

We noticed that a_t and a_{t-1} often differ greatly. In order to achieve the new portfolio $w_{t+1} = a_t$, the agents would need to make many different trades. Sometimes, the transactions only involve selling or buying a very small number of stocks.

To tackle the issue of having an overly high number of daily transactions, we introduced threshold k, which acts as a cut-off point in deciphering whether the

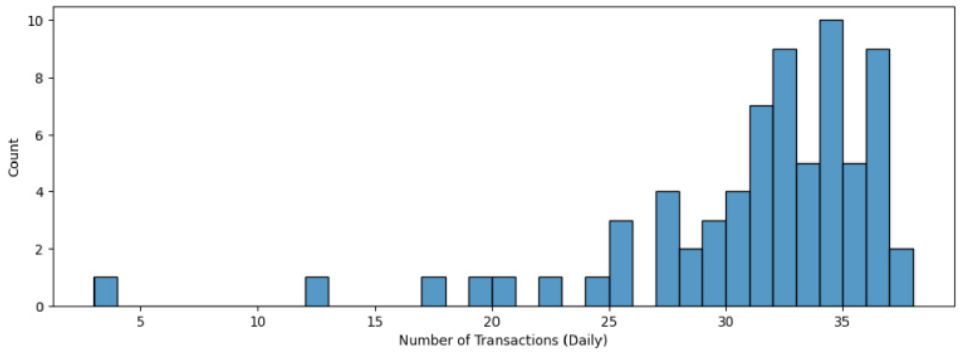

Figure 4. Distribution of number of transactions per day for PPO agent. Agent was averaging 30.67 transactions per day.

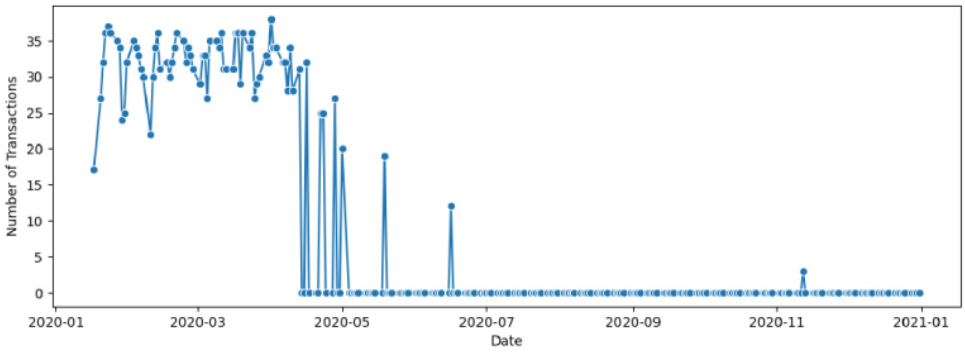

Figure 5. Trend of number of transactions per day for PPO agent. Large number of transactions at the start of the trading period resulting in the agent running out of money towards the end.

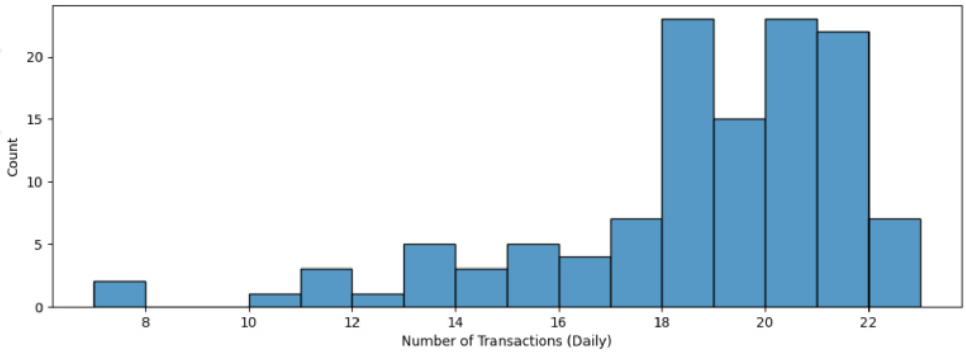

Figure 6. Distribution of number of transactions per day for A2C agent. Agent was averaging 18.27 transactions per day.

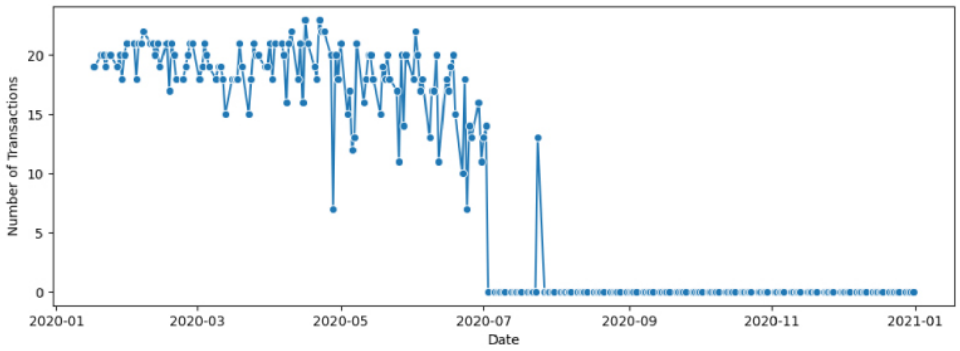

Figure 7. Trend of number of transactions per day for A2C agent. Similar trading pattern to the PPO agent resulting in A2C agent running out of money due to transaction cost.

trade should be exercised. The trade involving stock i was allowed only if

$$a_{t,i} - w_{t,i} > k. \tag{7}$$

Trades would only occur if the agent decided that the weight of a certain stock needs to change by at least k. To achieve an optimal threshold value for both agents, we conducted cross-validation across a range of values from 0.04 to 0.06 with an interval of 0.005. It was observed in Figures 8 and 9 that the performance of the PPO and A2C agents had improved.

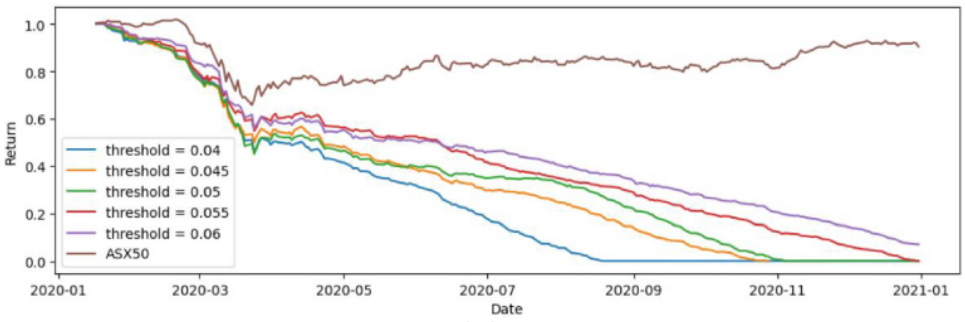

Figure 8. Cumulative returns of PPO agents with different transaction thresholds. PPO agents still made too many transactions, which resulted in the balance going to zero.

However, the results were still undesirable. The PPO agent continued to spiral out of control, while the A2C agent failed to beat the ASX50 market. We also noticed that the A2C agent (k = 0.06) did not make any transactions, which resulted in a stagnant trendline.

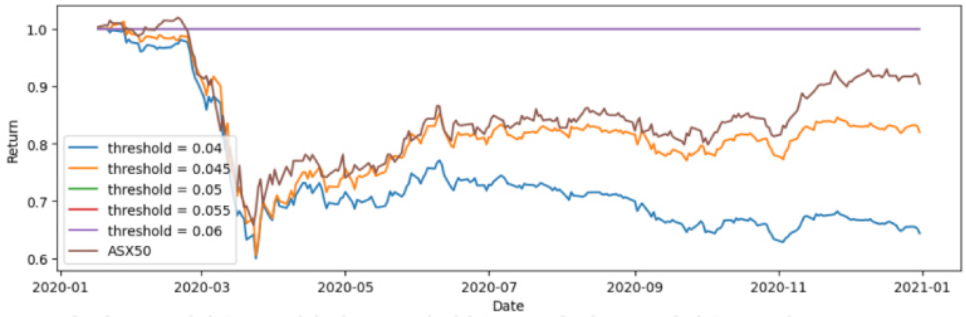

Figure 9. Cumulative returns of A2C agents with different transaction thresholds. The balance of the A2C agents did not go to zero but they were still underperforming.

5.1. High transaction frequency

We investigated the trading history of the A2C agent with the best results (k = 0.045). Figures 10 and 11 show the distribution and trend of number of transactions made per day, respectively. The A2C agent made around 2.47 transactions, paying a transaction cost of $49.40 per day. The number of transactions per day has decreased significantly. However, we observed that the frequency of transactions remained very high. We could increase the threshold further, but as shown in Figure 9, the result did not improve and eventually k will be too high for the agent to make any trades.

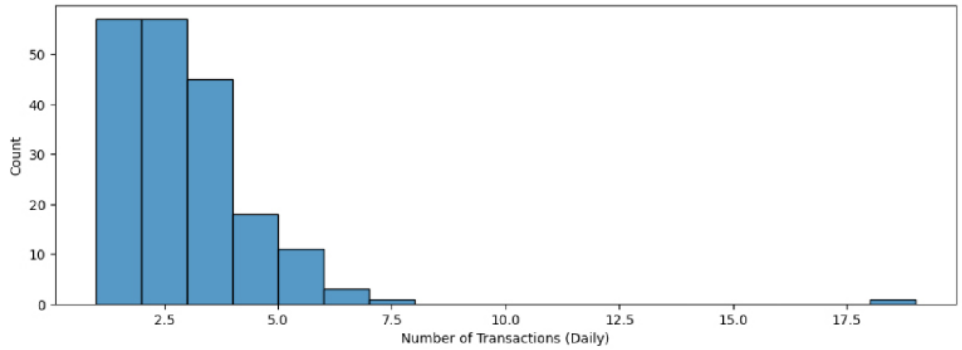

Figure 10. Distribution of number of transactions per day for A2C agent (k = 0.045). Transactions per day had gone down significantly.

We then looked into one of the holdings the A2C agent (k = 0.045) had. In Figure 12, we noticed that the agent was making some peculiar trades. The agent had the intention to hold CBA.ax for a very long term. However, once in a while, it will sell the entire holding before buying it back shortly after. Every such sell and buy would cost the agent $40, which was an unnecessary expense. The agent was very rash in making its decisions and did not have a long-term plan in mind.

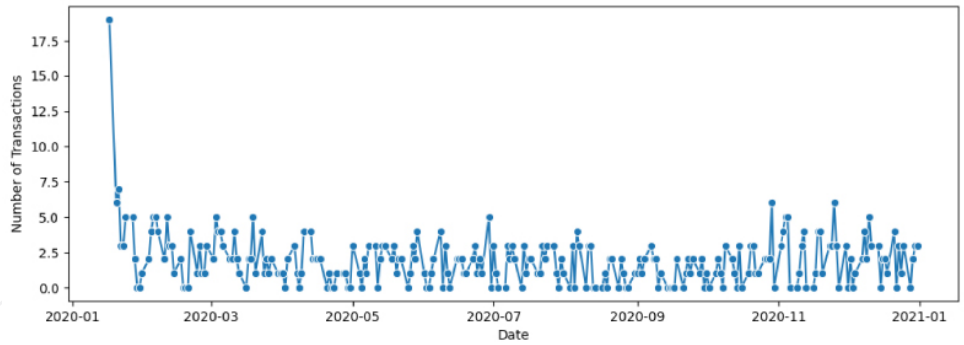

Figure 11. Trend of number of transactions per day for A2C agent (k = 0.045). The agent is transacting almost every single day, which is too frequent.

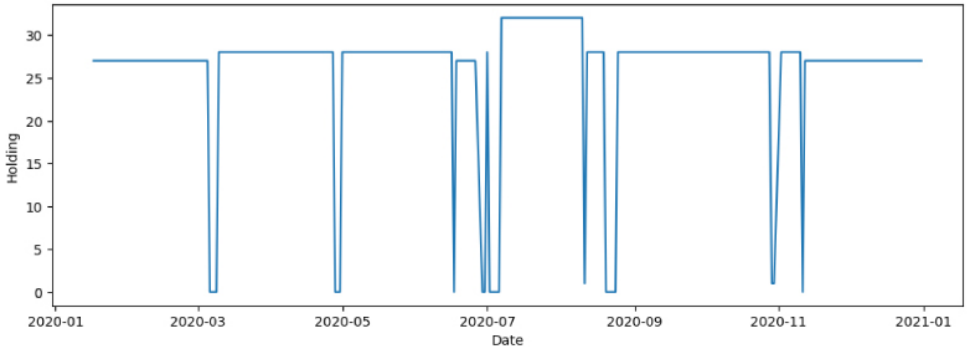

Figure 12. CBA holdings in the A2C portfolio (k = 0.045). There were many peculiar instances where the agent sells the whole portfolio before buying it all back.

Therefore, instead of simply comparing a_t against w_t, we computed a weighted moving average of a_t. The agent would only be able to execute a trade involving stock i if

$$\mathrm{WS}_w(t,i) = \sum_{j=0}^{w} \frac{(w-j) \times a_{t-j,i}}{w} \tag{8}$$

$$\mathrm{WMA}_w(t,i) = \frac{\mathrm{WS}_w(t,i)}{\sum_{j=1}^{n} \mathrm{WS}_w(t,j)} \tag{9}$$

$$\mathrm{WMA}_w(t,i) - w_{t,i} > k \tag{10}$$

where $\mathrm{WS}_w(t,i)$ is the weighted sum of a_t in a window w.

Trades would only occur if the agent decided that the weight of a certain stock needs to change by k for a time period w. To achieve an optimal threshold value for both agents, we conducted cross-validation across three values: w = 5, 10, and 15.

The results of the PPO agent improved substantially as shown in Figure 13. The agent is able to beat the index, with a few models achieving decent returns. The PPO agent (k = 0.04, w = 15) reported a cumulative return of 13.25% but an annual volatility of 15.56% as shown in Table 2. However, the market experienced a crash in the first quarter of 2020. A model like the PPO agent (k = 0.045, w = 15) is too conservative. It has a much lower annual volatility but the upside after the crash was not as high as the PPO agent (k = 0.04, w = 15). Therefore, we chose the PPO agent (k = 0.04, w = 15) as our final PPO model.

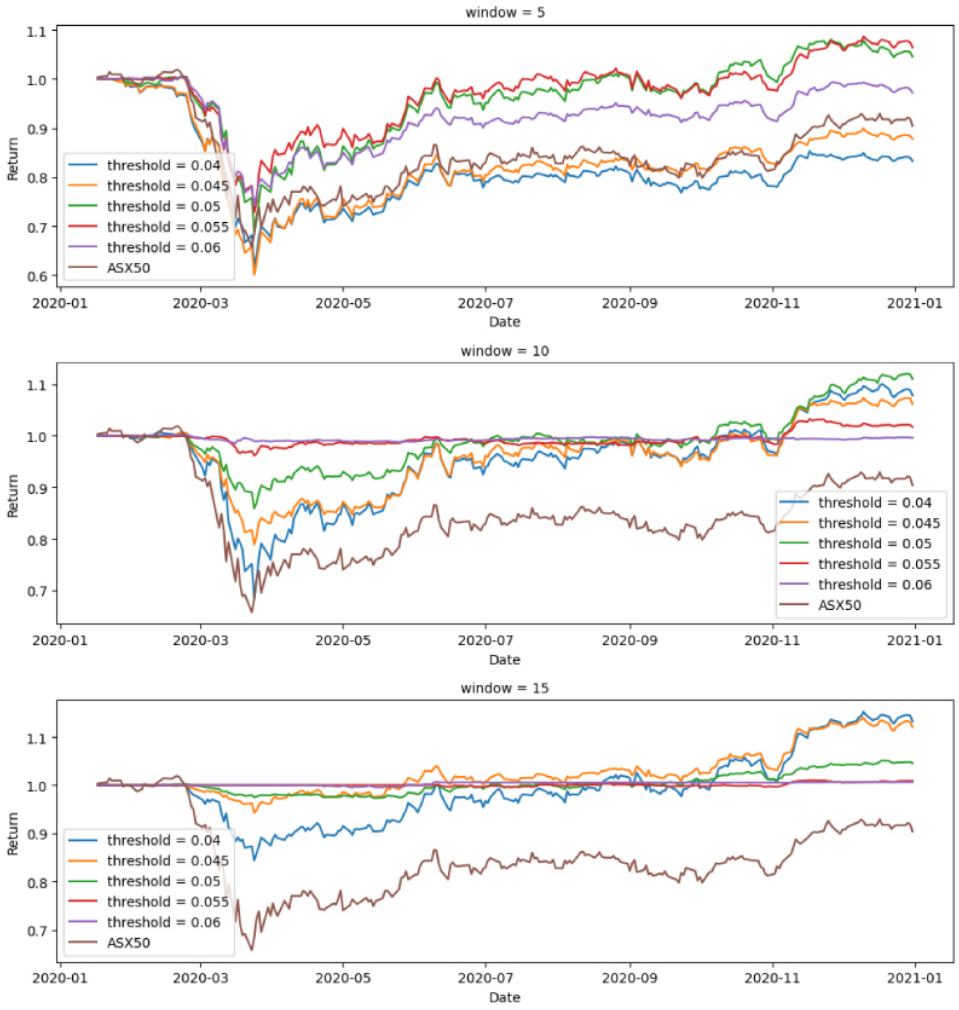

Figure 13. Cumulative returns for PPO agent with weighted moving average weights.

The results of the A2C agent improved substantially as shown in Figure 14. The agent is able to beat the index, with one model achieving remarkable returns. The A2C agent (k = 0.05, w = 15) reported a cumulative return of 23.68% but an annual volatility of 15.64% as shown in Table 3. Therefore, we chose the A2C agent (k = 0.05, w = 15) as our final A2C model.

Table 2. Summary of validation results for PPO agent with weighted moving average weights. The best PPO agent is k = 0.04, w = 15 with 13.25% validation CAGR.

	Transaction threshold				
Window size	**0.04**	**0.045**	**0.05**	**0.055**	**0.06**
5	−16.76%	−12.23%	4.56%	6.45%	−2.79%
	(25.29%)	(27.02%)	(24.44%)	(22.64%)	(17.79%)
10	7.83%	6.17%	11.05%	1.71%	−0.39%
	(25.77%)	(17.09%)	(12.14%)	(4.10%)	(1.49%)
15	**13.25%**	12.06%	4.56%	0.85%	0.61%
	(15.56%)	(9.04%)	(3.68%)	(1.37%)	(0.88%)

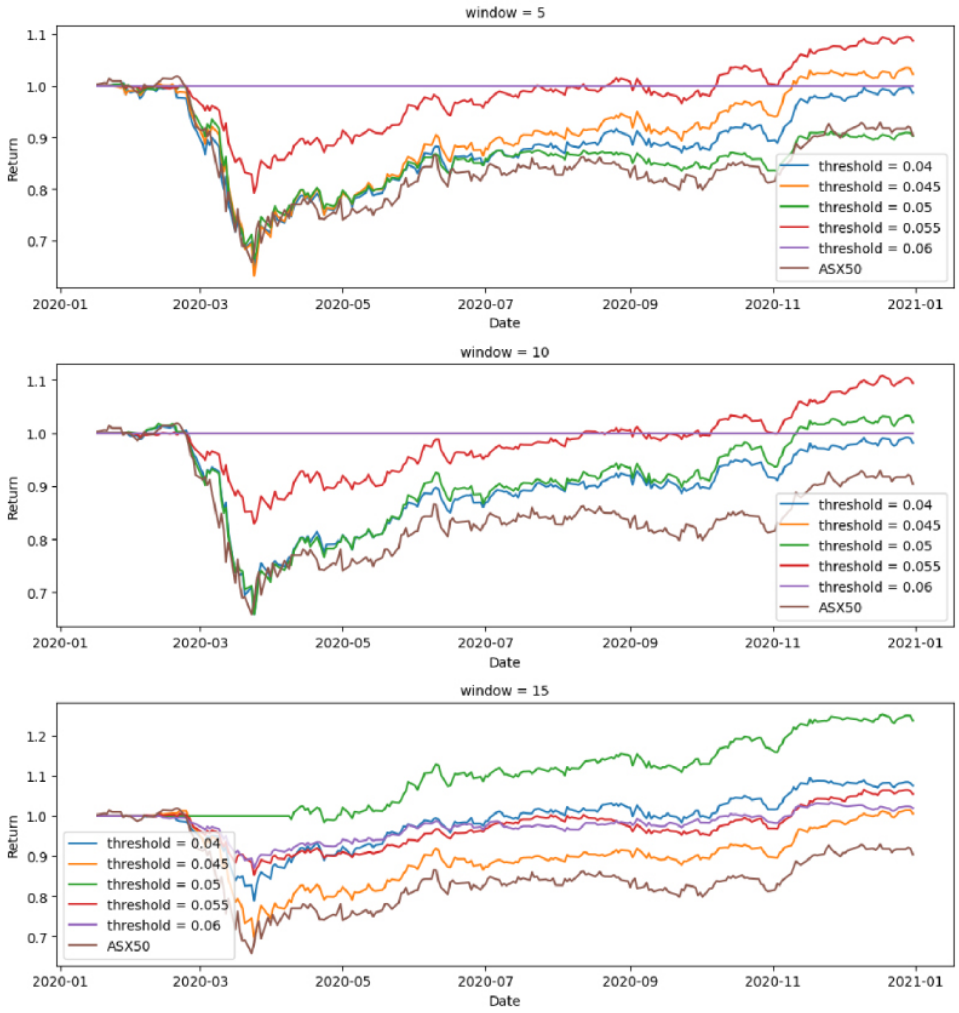

Figure 14. Cumulative returns for A2C agent with weighted moving average weights.

Table 3. Summary of validation results for A2C agent with weighted moving average weights.

Window size	Transaction threshold				
	0.04	0.045	0.05	0.055	0.06
5	−1.41%	2.27%	−9.71%	8.71%	0.0%
	(25.28%)	(24.97%)	(21.29%)	(18.8%)	(0.0%)
10	−1.82%	0.0%	2.04%	9.44%	0.0%
	(24.85%)	(0.0%)	(23.75%)	(17.1%)	(0.0%)
15	7.53%	0.58%	**23.68%**	5.46%	1.95%
	(17.14%)	(21.52%)	**(15.64%)**	(13.26%)	(11.28%)

5.2. Overall performance

Figure 15 illustrates the cumulative returns of the final PPO and A2C agents against the benchmark, ASX50 Index. The PPO agent was able to achieve a higher CAGR compared to the A2C agent. Both agents reported a positive cumulative return; however, both agents fell short of beating the index.

Figure 15. Cumulative returns of the final PPO and A2C agents against ASX50 Index.

Table 4 shows the summary statistics of the performance metrics by the agents and the ASX50 Index. Both the PPO (4.83%) and A2C (1.32%) had a lower cumulative return than the index (8.03%). The PPO portfolio had the lowest annual volatility (9.47%) while the A2C (13.98%) and the index (13.43%) had similar annual volatility. The A2C had the highest maximum drawdown (−24.59%) while the PPO had the lowest maximum drawdown (−11.76%). This shows that the PPO portfolio takes much lower risk as compared to the A2C model and was able to perform better. However, the SR of the PPO portfolio is only 0.29. The reduced level of risk taken did not produce sufficient returns.

However, if we examined each year of the testing period separately, we would get completely different results. Table 5 shows the summary statistics of the

Table 4. Summary statistics of final PPO and A2C agents against ASX50 Index.

Portfolio	CAGR (%)	Cumulative return (%)	Annual volatility (%)	SR	MDD (%)
PPO	2.34	4.83	9.47	0.29	−11.76
A2C	0.65	1.32	13.98	0.12	−24.59
ASX50	3.85	8.03	13.43	0.35	−14.59

Table 5. Summary statistics of final PPO and A2C agents against ASX50 Index (2021). Both agents performed well in a bullish market. A2C agent was able to generate twice the CAGR compared to PPO agent and the index.

Portfolio	CAGR (%)	Cumulative return (%)	Annual volatility (%)	Sharpe ratio	Maximum drawdown (%)
PPO	10.41	10.41	4.95	2.06	−4.67
A2C	21.78	21.78	9.38	2.18	−6.73
ASX50	10.38	10.38	11.16	0.95	−6.75

performance metrics by the agents and the ASX50 Index in 2021. This was a bull market, evident from the cumulative return of ASX50 (10.38%). Both the A2C (21.78%) and PPO (10.41%) portfolios had higher returns than the index. Additionally, both PPO (4.95%) and A2C (9.38%) recorded lower annual volatility than the index (11.15%). The SR was "very good" for both the PPO (2.06) and A2C (2.18) portfolios [33]. The PPO (−4.67%) portfolio was able to report the lowest maximum drawdown again.

Table 6 shows the summary statistics of the performance metrics by the agents and the ASX50 Index in 2022. This was a bear market, evident from the negative cumulative return of ASX50 (−2.14%). Both the A2C (−16.37%) and PPO (−4.70%) portfolios had lower returns than the index. The PPO portfolio was able to report the lowest maximum drawdown and annual volatility at −11.76% and 12.29%, respectively.

Table 6. Summary statistics of final PPO and A2C agents against ASX50 Index (2022). Both agents performed worse than the index in a bearish market. A2C suffered more compared to PPO agent.

Portfolio	CAGR (%)	Cumulative return (%)	Annual volatility (%)	Sharpe ratio	Maximum drawdown (%)
PPO	−4.70	−4.70	12.29	−0.31	−11.76
A2C	−16.37	−16.37	17.18	−0.91	−24.59
ASX50	−2.04	−2.14	15.30	−0.06	−14.46

5.2.1. PPO portfolio

Table 7 shows the top gains and top losses in the PPO portfolio. The PPO agent had made trades with 37 out of the 45 available stocks. The agent made profit in 21 of them with an average of $392.69 per stock and suffered a loss in 16 of them with an average of −$341.74 per stock. The agent made 66 (40 buys and 20 sells) transactions over the 2-year period paying $1320 in transaction cost.

Table 7. Top gains and losses of the PPO agent. PPO agent made the most from trading WTC with $1254.56 profit and lost the most with JHX with $1140.60 loss.

Ticker	Return	Ticker	Return	Ticker	Return	Ticker	Return
TLS	544.59	BXB	311.28	AIA	60.51	ALL	426.24
NCM	332.46	MQG	−458.77	DXS	−758.15	NST	631.95
CBA	138.43	REA	397.93	ANZ	−448.56	SEK	320.31
JHX	−1140.60	BSL	251.92	SCG	−176.51	RHC	618.76
IAG	−275.40	FMG	−272.60	SHL	453.75	WTC	1254.56
CSL	−11.97	TAH	229.50	TCL	−216.16	WOW	5.11
AMC	405.24	FPH	−67.54	GMG	285.66	COH	−373.68
WES	−423.84	QBE	237.80	REH	312.92	ASX	−699.67
RMD	447.95	SGP	−93.08	BHP	579.63	MGR	−44.63
NAB	−6.70						

Figure 16 shows the cumulative return of the top loser JHX against the ASX50 Index. The agent bought 60 JHX at $45.31 on 22/02/2022 and it is down 41.96% as of 06/01/2023. The agent seemed to believe the stock will eventually rebound and still held on to the stock. Figure 17 shows the cumulative return of the top winner WTC against the ASX50 Index. The agent bought 85 WTC at $31.01 on 2021/07/05. The stock price surges by 47.60% over the next 2 months. The agent then sold the stock at $45.77 on 2021/09/01. The agent had not traded the stock again since.

Figure 16. Cumulative return of the top loser (JHX) against ASX50 Index.

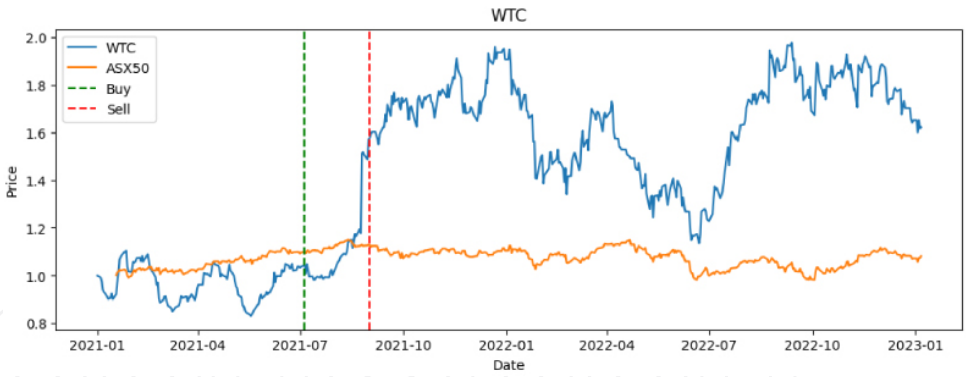

Figure 17. Cumulative return of the top winner (WTC) against ASX50 Index.

5.2.2. A2C portfolio

Table 8 shows the top gains and top losses in the A2C portfolio. The A2C agent had made trades with 17 out of the 45 available stocks. The agent made profit in eight of them with an average of $826.03 per stock and suffered a loss in nine of them with an average of −$656.27 per stock. The agent made 19 transactions (17 buys and 2 sells) over the 2-year period paying $380 in transaction cost. As compared to the PPO agent, the A2C agent keeps a smaller portfolio and tends to hold on to a given stock for a longer period of time.

Table 8. Top gains and losses of the A2C agent. A2C agent made the most from trading S32 with $2067.10 profit and lost the most with TAH with $2484.12 loss.

Ticker	Return	Ticker	Return	Ticker	Return	Ticker	Return
BHP	145.59	QBE	1900.92	REH	−345.87	SHL	−375.7
COL	−246.13	MGR	−299.78	BSL	−84.80	RMD	642.40
SUN	564.00	TLS	685.44	XRO	−1301.74	CBA	−98.01
S32	2067.10	SCG	389.12	TAH	−2484.12	GMG	−670.24
RIO	213.64						

Figure 18 shows the cumulative return of the top loser TAH against the ASX50 Index. The agent bought 652 TAH at $4.90 on 08/06/2022. The TAH share price crashed significantly, dropping by a massive 82%. This hurt the returns of the A2C portfolio really badly and is a major reason for the A2C portfolio's bad performance in 2022. Figure 19 shows the cumulative return of the top winner S32 against the ASX50 Index. The agent bought 1090 S32 at $2.72 on 2021/02/23. The agent held on to the stock for about a year and then sold 630 (58%) of the stock at $4.96 (82.35% increase) on 2022/04/13. The agent held on to the remaining 430 stocks since.

The A2C agent has the tendency to buy a stock and hold it for a very long term. In fact, it had only made two sells over the past 2 years but have built up a portfolio of

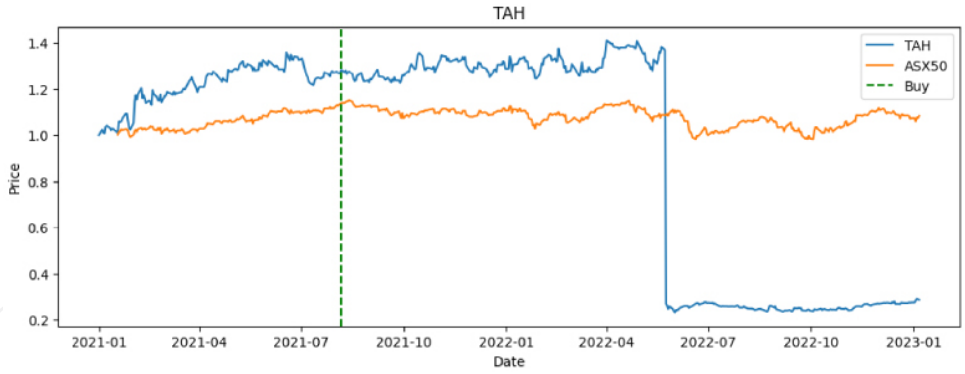

Figure 18. Cumulative return of the top loser (TAH) against ASX50 Index.

Figure 19. Cumulative return of the top winner (S32) against ASX50 Index.

17 stocks. One was S32.ax as shown in Figure 19. The other sell was TLS as shown in Figure 20. The threshold value that was determined by the validation process might not be suitable for the market conditions in the testing periods, resulting in critical trades not executed.

Figure 20. Cumulative return of TLS against ASX50 Index.

6. Conclusion

In conclusion, we applied two DRL algorithms with continuous action space to tackle the portfolio management problem in the Australian market. In order to reduce the frequency and volume of trading made by the agents, we introduced a weighted moving average and a transaction threshold to determine whether the trade executed by the agents is necessary. The weighted moving average and threshold were able to help reduce the number of trivia trades made by the agent. The optimal windows for the moving average and the optimal threshold were determined to be 15 and 0.04 for PPO and 15 and 0.05 for A2C, respectively.

The trading agents were unable to outperform the ASX50 Index, but they were still able to somewhat capture the patterns of the market movement. The A2C agent was better at following trends and have the higher upside potential but can suffer from more severe damage during bearish markets. On the other hand, the PPO agent has the lowest annual volatility and the highest maximum drawdown, which is more helpful in a bearish or volatile market. The opposite conclusion was drawn in the study by Yang *et al.* In their study, they observed that the A2C agent had the lowest annual volatility and the highest maximum drawdown while the PPO agent acts better in generating more returns [24].

The DRL still has many limitations when being applied to portfolio management. Some of the limitations include the following:

1. Data limitations: The DRL requires a large amount of high-quality data for the agents to be trained effectively. However, high-quality financial data is not easily accessible to the average person. The OHLCV data is very noisy and is influenced by many external variables that are not predictable. This makes it very challenging for one to develop an agent that can respond to any market conditions.
2. Overfitting: The DRL can be prone to overfitting, which can be evident in this study. The agents were able to trade and perform well in the validation period as well as the first half of the testing period. However, when faced with a situation, they will make incorrect investment decisions resulting in poor performance.
3. Black-box models: It is very difficult to interpret how the agents are able to come up with those decisions. In our study, we were able to identify the top losers of the portfolio, but we were unable to say for sure why the agent made such a bad trade. We could only speculate based on the retrospective information. This can make it hard to explain the agent's decisions to investors, which is often more important than the results itself.

4. Real-world implementation: The simulated stock environment is built upon many assumptions and cannot fully cover the complexity of the stock market in the real world. There are many other hurdles that need to be overcome.

For future works, there is a limitation of using a fixed transaction threshold when deciding on which trades are necessary and which are not. It will be interesting to explore a more dynamic way to determine such a threshold. Second, many researchers have proposed including market sentiment into the model, which we agree is a crucial factor to stock trading. However, financial natural language processing (NLP) is non-trivial and a separate field of research on its own. It will be interesting to see if any breakthrough in the field of NLP will bring new perspectives to DRL in portfolio management. Lastly, an explainable artificial intelligence (XAI) set of frameworks can help understand and interpret the decisions made by the agents. It will be interesting to explore improving the explainability of DRL agents in portfolio management, which could give us new insights into how we approach trading stocks.

CRediT authorship contribution statement

Weiye Wu: Conception and design of study, Analysis and interpretation of data, Writing of original draft.

Carol Anne Hargreaves: Acquisition of data, Supervision of study, Writing – review & editing.

Declaration of competing interests

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

Data availability

Data will be made available on request.

Funding

This research received no external funding.

Appendix A. List of 45 constituent stocks of ASX50 used in this study

SN	Ticker	Company
1	AIA	Auckland International Airport Ltd
2	ALL	Aristocrat Leisure Ltd
3	AMC	Amcor Plc
4	ANZ	Australia and New Zealand Banking Group Ltd
5	APA	APA Group
6	ASX	ASX Ltd
7	BHP	BHP Group Ltd
8	BSL	Bluescope Steel Ltd
9	BXB	Brambles Ltd
10	CBA	Commonwealth Bank of Australia
11	COH	Cochlear Ltd
12	COL	Coles Group Ltd
13	CSL	CSL Ltd
14	DXS	Dexus
15	FMG	Fortescue Metals Group Ltd
16	FPH	Fisher & Paykel Healthcare Corporation Ltd
17	GMG	Goodman Group
18	IAG	Insurance Australia Group Ltd
19	JHX	James Hardie Industries Plc
20	MGR	Mirvac Group
21	MQG	Macquarie Group Ltd
22	NAB	National Australia Bank Ltd
23	NCM	Newcrest Mining Ltd

24	NST	Northern Star Resources Ltd
25	QBE	QBE Insurance Group Ltd
26	REA	REA Group Ltd
27	REH	Reece Ltd
28	RHC	Ramsay Health Care Ltd
29	RIO	RIO Tinto Ltd
30	RMD	Resmed Inc
31	S32	SOUTH32 Ltd
32	SCG	Scentre Group
33	SEK	Seek Ltd
34	SGP	Stockland
35	SHL	Sonic Healthcare Ltd
36	STO	Santos Ltd
37	SUN	Suncorp Group Ltd
38	TAH	Tabcorp Holdings Ltd
39	TCL	Transurban Group
40	TLS	Telstra Corporation Ltd
41	WBC	Westpac Banking Corporation
42	WES	Wesfarmers Ltd
43	WOW	Woolworths Group Ltd
44	WTC	Wisetech Global Ltd
45	XRO	Xero Ltd

Appendix B. Details of technical indicators

EMA is a trend-following indicator that places a greater emphasis on recent data points. As compared to a simple moving average, EMA is more sensitive to recent price movements, and thus is more reliable and relevant. EMA can be expressed

as follows:

$$\mathrm{EMA}_n(t) = \frac{2 \times c_t + (n-1)\mathrm{EMA}_n(t-1)}{n+1} \tag{B.1}$$

where c_t is the close price at time t and n is the time span.

MACD is a momentum indicator that shows the relationship between two EMAs, calculated by taking the difference between a long-term EMA and a short-term EMA. MACD can be expressed as follows:

$$\mathrm{MACD}(t) = \mathrm{EMA}_{12}(t) - \mathrm{EMA}_{26}(t). \tag{B.2}$$

RSI is a momentum indicator that measures the speed and magnitude of the stock price movement to determine whether the stock is overbought or oversold by comparing the average gains and loss over a specified time period. RSI can be expressed as follows:

$$\mathrm{RSI} = 100 - \frac{100}{1 + \mathrm{RS}} \tag{B.3}$$

where RS is the average gains over average loss over a period.

Stochastic oscillator is a momentum indicator that compares the close price to its price range over a specific period of time. The stochastic oscillator can be expressed as follows:

$$\%K = \frac{c_t - L}{H - L} \times 100 \tag{B.4}$$

where L and H are the lowest and highest prices over a period.

OBV is a volume indicator that measures the buying and selling pressure of the stock, using the trading volume movement to predict the stock price movement. OBV can be expressed as follows:

$$\mathrm{OBV}_t = \begin{cases} \mathrm{OBV}_{t-1} + v_t, & \text{if } c_t > c_{t-1} \\ \mathrm{OBV}_{t-1} - v_t, & \text{if } c_t < c_{t-1} \\ \mathrm{OBV}_{t-1}, & \text{otherwise} \end{cases} \tag{B.5}$$

where v_t is the trading volume at time t.

VPT is a volume indicator that measures the strength of a trend based on the relationship between the trading volume movement and stock price movement. VPT

can be expressed as follows:

$$\mathrm{VPT}_t = \mathrm{VPT}_{t-1} + v_t \times \frac{c_t - c_{t-1}}{c_{t-1}}. \tag{B.6}$$

MFI is a volume indicator that measures the inflow and outflow of money into a stock over a specific period of time. Similar to RSI, MFI is used to identify overbought and oversold signals, but it incorporates both price and volume data. MFI can be expressed as follows:

$$\mathrm{MFI} = 100 - \frac{100}{1 + MFR} \tag{B.7}$$

$$\mathrm{TP}_t = \frac{h_t + l_t + c_t}{3} \times v_t \tag{B.8}$$

where *MFR* is the average typical price (TP) gains over average loss over a period.

Bollinger Band is a volatility indicator that is defined by two trendlines, each at two standard deviations away from a simple moving average (one above and one below). The Bollinger Band can be expressed as follows:

$$\mathrm{BB} = \mathrm{SMA}_n \pm \sigma_n \tag{B.9}$$

where $\mathrm{SMA}_n(t)$ is the simple moving average and $\sigma_n(t)$ is the standard deviation of TP over the last n periods.

References

1 Saul D. *Retail trading just hit an all-time high. here's what stocks are the most popular [Internet]*. Forbes; 2023 Feb 6 [cited 2023 Mar 2]. Available from https://www.forbes.com/sites/dereksaul/2023/02/03/retail-trading-just-hit-an-all-time-high-heres-what-stocks-are-the-most-popular/.

2 Chan EP. *Quantitative trading: how to build your own algorithmic trading business*. New Jersey: Wiley; 2021.

3 Li Y, Zheng W, Zheng Z. Deep robust reinforcement learning for practical algorithmic trading. *IEEE Access*. 2019;**7**: 108014–108022. doi:10.1109/access.2019.2932789.

4 Jing N, Wu Z, Wang H. A hybrid model integrating deep learning with investor sentiment analysis for stock price prediction. *Expert Syst Appl*. 2021;**178**: 115019. doi:10.1016/j.eswa.2021.115019.

5 Rezaei H, Faaljou H, Mansourfar G. Stock price prediction using deep learning and frequency decomposition. *Expert Syst Appl*. 2021;**169**: 114332. doi:10.1016/j.eswa.2020.114332.

6 Agrawal M, Kumar Shukla P, Nair R, Nayyar A, Masud M. Stock prediction based on technical indicators using deep learning model. *Comput Mater Contin*. 2022;**70**(1):287–304. doi:10.32604/cmc.2022.014637.

7 Li Y, Ni P, Chang V. Application of deep reinforcement learning in stock trading strategies and stock forecasting. *Computing*. 2019;**102**(6):1305–1322. doi:10.1007/s00607-019-00773-w.

8 Henderson P, Islam R, Bachman P, Pineau J, Precup D, Meger D. Deep reinforcement learning that matters. *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32, 2018 doi:10.1609/aaai.v32i1.11694.

9 Ali Imran Z, Wong W-C, Ismail R. Momentum effect all over the world. *Int J Bank Finance*. 2020;**14**: 75–93. doi:10.32890/ijbf2019.14.0.9912.

10 Yue H, Liu J, Tian D, Zhang Q. A novel anti-risk method for portfolio trading using deep reinforcement learning. *Electronics*. 2022;**11**(9):1506. doi:10.3390/electronics11091506.

11 Wang Z, Huang B, Tu S, Zhang K, Xu L. DeepTrader: A deep reinforcement learning approach for risk-return balanced portfolio management with market conditions embedding. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. vol. 35, 2021. p. 643–650. doi:10.1609/aaai.v35i1.16144.

12 Théate T, Ernst D. An application of deep reinforcement learning to algorithmic trading. *Expert Syst Appl*. 2021;**173**: 114632. doi:10.1016/j.eswa.2021.114632.

13 Markowitz H, Todd GP. *Mean-variance analysis in portfolio choice and capital markets*. New Jersey: Wiley; 2000.

14 Rubinstein M. Markowitz's "Portfolio selection": a fifty-year retrospective. *J Finance*. 2002;**57**(3):1041–1045. doi:10.1111/1540-6261.00453.

15 Huang S-H, Miao Y-H, Hsiao Y-T. Novel deep reinforcement algorithm with adaptive sampling strategy for continuous portfolio optimization. *IEEE Access*. 2021;**9**: 77371–77385. doi:10.1109/access.2021.3082186.

16 Hawley JP, Beyhaghi M. Modern portfolio theory and risk management: assumptions and unintended consequences. *SSRN Electron J*. 2011;**3**(1):17–37. doi:10.2139/ssrn.1923774.

17 Brini A, Tantari D. Deep reinforcement trading with predictable returns. *Physica A*. 2023;**622**: 128901. doi:10.1016/j.physa.2023.128901.

18 Chaoki A, Hardiman S, Schmidt C, Serie E, De Lataillade J. Deep deterministic portfolio optimization [Internet]. *J Finan Data Sci*. 2020;**6**: 16–30. Science Direct. Available from https://www.sciencedirect.com/science/article/pii/S2405918820300118.

19 Chen L, Gao Q. Application of deep reinforcement learning on automated stock trading. In: *2019 IEEE 10th International Conference on Software Engineering and Service Science (ICSESS)*. Beijing, China: IEEE; 2019. p. 29–33. doi:10.1109/icsess47205.2019.9040728.

20 Dang Q-V. *Reinforcement learning in stock trading. Advanced computational methods for knowledge engineering*. Berlin: Springer, CHAM; 2019. p. 311–322. doi:10.1007/978-3-030-38364-0_28.

21 Li Y, Yang X, Li F, Zhou P. An improved reinforcement learning model based on sentiment [Internet]. Paper; 2021 Feb 2 [cited 2023 Mar 2]. Available from https://ideas.repec.org/p/arx/papers/2111.15354.html.

22 Brim A, Flann NS. Deep reinforcement learning stock market trading, utilizing a CNN with candlestick images. *PLoS One*. 2022;**17**(2):e0263181. https://doi.org/10.1371/journal.pone.0263181.

23 Liang Z, Chen H, Zhu J, Jiang K, Li Y. Adversarial deep reinforcement learning in portfolio management [Internet]. arXiv; 2018 Nov 18 [cited 2023 Mar 2]. Available from https://arxiv.org/abs/1808.09940.

24 Yang H, Liu X-Y, Zhong S, Walid A. Deep reinforcement learning for automated stock trading: an ensemble strategy. *SSRN Electron J*. 2020; doi:10.2139/ssrn.3690996.

25 Sadriu L. Deep reinforcement learning approach to portfolio optimization[Internet]; 2022. Available from http://lup.lub.lu.se/student-papers/record/9071680.

26 Koratamaddi P, Wadhwani K, Gupta M, Sanjeevi SG. Market sentiment-aware deep reinforcement learning approach for stock portfolio allocation. *Eng Sci Technol Int J*. 2021;**24**(4):848–859. doi:10.1016/j.jestch.2021.01.007.

27 Zhang Z, Zohren S, Roberts S. Deep reinforcement learning for trading. *J Financ Data Sci*. 2020;**2**(2):25–40. doi:10.3905/jfds.2020.1.030.

28 ASX 50 list. Constituents, sectors & weighting. (n.d.) [Internet]; [cited 2023 Mar 2]. Available from https://www.asx50list.com/.

29 Neely CJ, Zhou G, Rapach DE, Tu J. Forecasting the equity risk premium: the role of technical indicators. Federal Reserve Bank of St. Louis Working Paper 2010-008. 2010. doi:10.20955/wp.2010.008.

30 Brockman G, Cheung V, Pettersson L, Schneider J, Schulman J, Tang J, et al. OpenAI Gym [Internet]. arXiv; 2016. Available from: https://arxiv.org/abs/1606.01540.

31 Shihao G, Bryan K, Dacheng X. Empirical asset pricing via machine learning. *Rev Finan Stud*. May 2020;**33**(5):2223–2273. doi:10.1093/rfs/hhaa009.

32 Raffin A, Hill A, Gleave A, Kanervisto A, Ernestus M, Dormann N. Stable-baselines3: Reliable reinforcement learning implementations [Internet]. *J Mach Learn Res*. 2021 Jan [cited 2023 Mar 2];**22**(1):12348–12355. Available from https://dl.acm.org/doi/abs/10.5555/3546258.3546526.

33 Baldridge R. *Understanding the Sharpe ratio* [Internet]. Forbes; 2022 Dec 14 [cited 2023 Mar 03]. Available from https://www.forbes.com/advisor/investing/sharpe-ratio/.