Neural Computing and Applications (2022) 34:7125–7139
https://doi.org/10.1007/s00521-021-06853-3

ORIGINAL ARTICLE

Dynamic portfolio rebalancing through reinforcement learning

Qing Yang Eddy Lim[1] · Qi Cao[2] · Chai Quek[1]

Received: 24 June 2021 / Accepted: 12 December 2021 / Published online: 27 December 2021

Abstract

Portfolio managements in financial markets involve risk management strategies and opportunistic responses to individual trading behaviours. Optimal portfolios constructed aim to have a minimal risk with highest accompanying investment returns, regardless of market conditions. This paper focuses on providing an alternative view in maximising portfolio returns using Reinforcement Learning (RL) by considering dynamic risks appropriate to market conditions through dynamic portfolio rebalancing. The proposed algorithm is able to improve portfolio management by introducing the dynamic rebalancing of portfolios with vigorous risk through an RL agent. This is done while accounting for market conditions, asset diversifications, risk and returns in the global financial market. Studies have been performed in this paper to explore four types of methods with variations in fully portfolio rebalancing and gradual portfolio rebalancing, which combine with and without the use of the Long Short-Term Memory (LSTM) model to predict stock prices for adjusting the technical indicator centring. Performances of the four methods have been evaluated and compared using three constructed financial portfolios, including one portfolio with global market index assets with different risk levels, and two portfolios with uncorrelated stock assets from different sectors and risk levels. Observed from the experiment results, the proposed RL agent for gradual portfolio rebalancing with the LSTM model on price prediction outperforms the other three methods, as well as returns of individual assets in these three portfolios. The improvements of the returns using the RL agent for gradual rebalancing with prediction model are achieved at about 27.9–93.4% over those of the full rebalancing without prediction model. It has demonstrated the ability to dynamically adjust portfolio compositions according to the market trends, risks and returns of the global indices and stock assets.

Keywords Reinforcement learning · Dynamic portfolio rebalancing · Portfolio optimisation · Price prediction

1 Introduction

In modern portfolio theory, portfolio optimisation is one of the objectives to maximise returns of a portfolio while minimising risks using diversification methods [1, 2]. Financial market risk analysis and behavioural risk studies are involved in optimising portfolios [3–5]. There are two common strategies for financial asset allocations within a portfolio to manage market risk [6]. One is strategic asset allocation (SAA) that attempts to balance risks and returns with different weightages for target asset allocations. The other is tactical asset allocation (TAA), that attempts to switch portfolio allocations to the most attractive asset proportions when certain financial markets trend changes are predicted using market forecasting tools [7]. In practice, a combination of SAA and TAA strategies is deployed in tandem to maximise their advantages by Asset Management firms (AMF), such as JPMorgan and Goldman Sachs. By combining SAA and TAA strategies, the asset allocation consists of a fixed percentage amount and a variable percentage amount in a portfolio depending on market conditions.

In the behavioural risk management, studies have been conducted to investigate how returns of a portfolio may be

✉ Qi Cao
qi.cao@glasgow.ac.uk

Qing Yang Eddy Lim
eddy0006@e.ntu.edu.sg

Chai Quek
ashcquek@ntu.edu.sg

[1] School of Computer Science and Engineering, Nanyang Technological University, Singapore, Singapore

[2] School of Computing Science, University of Glasgow, Singapore Campus, Singapore, Singapore

affected by the risk adversity of individuals, such as portfolio managers in AMF [8]. The risk adversity of an individual can range in a risk spectrum being from risk averse at one end, to risk seeking at the other end with varying degrees of loss aversion and sensitivity [9]. According to prior research outcomes on the loss aversion, people usually are more sensitive to losses than gains [10]. It will affect individual decision-makings and portfolio asset prices in financial markets [11, 12]. As such, the risk adversity of individuals may affect the financial market volatility.

It is beneficial to explore techniques in artificial intelligence (AI) and machine learning algorithms as portfolio construction strategies in fund management involving SAA and TAA approaches. AI and machine learning algorithms have been utilised to maximise the returns of constructed portfolios with self-learning and less human interventions [13], such as evolutionary computation [14–16], genetic algorithms (GA) [14, 17, 18], particle swarm optimization algorithm [19, 20], fuzzy logic [21], reinforcement learning (RL) [22, 23], and recurrent reinforcement learning (RRL) [1]. Fuzzy neural network is also used for market risk prediction [24], with such information being useful for portfolio construction. Serrano [25] presents the research work using random neural network (RNN) to predict stock market index prices.

RL is a type of machine learning algorithm being used in various applications. The RL agent has learning capability through interaction with its environment [26]. An action is decided by the RL agent, according to the current state in the environment. The action is going to change the current state into the next state. A reward is given to the RL agent for each action. A new action will be decided by the RL agent in the new state of the environment [27]. The iteration repeats for the RL agent, aiming to achieve maximised total rewards. In some scenarios, multiple agents can learn and work collaboratively to change the environment to suit certain needs [28].

RL is reported to solve problem statements of financial industry, such as pricing strategy optimization in insurance industry [29], bank marketing campaigns offering credit card services [30], and portfolio managements [31]. RL has been utilised for trading of financial assets on the stock and foreign exchange market. Almahdi and Yang [32] introduce RRL-based portfolio management method for computing and optimizing investment decisions with time efficiency by incorporating past investments actions in time-stacks. The experiments have been conducted for a portfolio with ten stocks selected from different sectors of S&P 500 in time frame of January 2013 to July 2017. Deep reinforcement learning (DRL) models with adjustable trading policy and stock performance indicator data have been presented for active portfolio management [33]. A portfolio management system is depicted using RL with multiple agents, each of which trades its own sub-portfolio under different policy in current market states [34]. The different actions of the multi-agent and diversified portfolios aim to diversify the risks and maximize the rewards of each agent. RL agents with two policy algorithms with Q-function for four states and five actions are reported [35], to re-allocate portfolio with two assets; one asset as S&P 500 Exchange Traded Funds (ETF), and the other as Barclays Capital U.S. Aggregate Bond Index (AGG) or a 10-year U.S. T-note. The discrete RL agent's actions to re-balance the two assets in the portfolio are taken quarterly, semi-annual or yearly without considering transaction costs and taxes. The performance of annual rebalancing frequency is observed to have better investment returns comparing to quarterly and semi-annual rebalancing frequency. A DRL and rule-based policy approach is presented for different versions of agents trading against each other in a continuous virtual environment [36]. The signals of relationships between actions and market behaviours are generated by the risk curiosity-driven learning to improve the quality of actions. Its performance and profitability are analysed through experiments using eight financial assets individually. Park et al. [37] derive a DRL trading strategy for multiple assets with experiments performed using two portfolios: one consisting of three EFT assets from the U.S. stock market, the other with three EFT assets from the Korean stock market. It reports that discrete action space modelling has more positive effects over continuous action space modelling in terms of lower turnover rate and more practical in real-world applications. The DRL model is utilized for algorithmic trading through learning from features of environmental states and financial signal representations to improve action decision-making [38]. The weights of the features including current financial product features, related financial products features and technical indicators are adjusted and re-assigned based on the learning outcomes to enhance the accuracy. A framework and trading agent implemented by DRL are employed for algorithmic trading with generic action set to adjust trading rules by learning the market conditions [39]. The effectiveness of the framework is evaluated through individual experiments on three stock and two index assets separately.

Many of the reported methods have certain assumptions, such as without considering transaction costs, where the profitability of the algorithm will be significantly impacted by transaction costs in practical scenarios [40]. Most research in RL for stock trading normally predicts trading strategies and decisions with trading a fixed number of shares, according to various trading signals, trends, features, and market conditions [41]. As another challenge in RL trading research, it is usually more difficult to predict varied number of shares for trading in actions under different market conditions. A RL algorithm with continuous-time, discrete-state for policy optimization has been introduced for managing financial portfolio, which is characterized by transaction costs involving time penalization [42]. Portfolio rebalancing is performed through performance measurement using the RRL method and adjusted objective function with the consideration of transaction costs and coherent risk of market conditions [1]. The buy or sell signals are generated and asset allocation weights are adjusted according to market volatility situations. Actions to sell stocks and stop-loss strategy are taken when the market volatility is high, while actions to buy stocks are taken by new generated re-enter signals. A portfolio consisting of five ETF assets in the time frame of January 2011 to December 2015 is selected for the experiments [1]. Jeong and Kim [41] combine RL and a deep neural network (DNN) for prediction by adding DNN regressor to a deep Q-network, with experiments conducted on four different stock indices individually: S&P500, KOSPI, HSI, and EuroStoxx50. It enables the predictions with different number of shares for each asset for the first time, compared to trading with fixed number of shares of other methods, that increases the trading profits. A deep Q-learning framework is introduced for portfolio management consisting of a global agent and multiple local agents, each of which handles trading of a single asset in the portfolio [43]. The global agent manages the rewards function for each local agent. The experiments are performed using a crypto portfolio consisting of four cryptocurrencies assets: Bitcoin, Litecoin, Ethereum and Riple in time frame from July 2017 to December 2018. The RL approach is reported to combine with the RNN to simulate investment decisions of asset bankers on profit marking on specific asset markets under different variables and configurations [44].

This paper focuses on providing an alternative view in maximising portfolio returns using RL by considering dynamic risks appropriate to market conditions and transaction costs through portfolio rebalancing. The proposed RL agent aims to improve returns of the portfolio Net Asset Value (NAV), by exploring four methods using a combination of full portfolio rebalancing, gradual portfolio rebalancing, without price prediction model, and with Long Short-Term Memory (LSTM) price prediction models. These four approaches will be presented and compared using three constructed financial portfolios. One of the portfolios consist of three global market indices with different risk levels. The other two portfolios consist of stock assets from different sectors of NYSE and NASDAQ markets with the presences of mixed market trends including bullish, bearish, and stagnant conditions. These assets in the portfolios are uncorrelated as much as possible. The performances of these four RL approaches for portfolio rebalancing will be discussed in this paper. Insights from this research will help portfolio managers to systematically improve the performance of portfolios by dynamically rebalancing asset allocations in tandem with changing market trends.

The main contributions of this paper are shown as follows:

- The portfolio rebalancing is performed by considering asset diversifications, risks, and returns using the combination of SAA and TAA strategies. The investment allocations to each asset in the portfolio are dynamically adjusted by the RL agent at run time.
- Market information lags are usually caused by lagging technical indicators that are computed using historical price data. The impacts of such lags in market trend detection have been analysed and mitigated using LSTM price prediction models in portfolio rebalancing methods.
- In the event of wrong action via portfolio rebalancing predicted by the proposed RL agents, the impacts of transaction/commission fees are analysed and compared using two different methods: full portfolio rebalancing and gradual portfolio rebalancing.

The remaining parts of the paper are organised as follows: Sect. 2 describes the proposed RL portfolio rebalancing methodology. Section 3 presents the proposed RL agent for dynamic portfolio rebalancing and the corresponding experiments. Section 4 concludes the paper.

2 Configurations of RL

Decision-makings of RL agents are based on Q values. A RL agent aims to determine a policy π and maximize the long-term rewards through a series of actions interacting with its environment, as shown in Eq. (1).

$$R_t = \sum_{k=0}^{T} \gamma^k r_{t+k+1} \tag{1}$$

where R_t is the accumulated sum of rewards till to the terminal time T; γ is the discount rate in the range of [0, 1]; and r_t is the reward received at the time t.

For the action a_t in the state s_t at the time t, the Q-value under a policy π (i.e. the value of a state-action pair) is derived by the expected return correspondingly [41], which is represented in Eq. (2).

$$\begin{aligned} Q^{\pi}(s,a) &= E[R_t | s_t = s, a_t = a] \\ &= E\left[\sum_{k=0}^{T} \gamma^k r_{t+k+1} | s_t = s, a_t = a\right] \end{aligned} \tag{2}$$

where the notation of $Q^{\pi}(s,a)$ is known as the action value function, i.e. the *Q-function* [41, 45].

The set-up of the RL agent defined in this paper is presented in this section.

2.1 Observable period

The observable period used by the proposed RL agent is the market prices in the range of 2014–2018. Actions of portfolio rebalancing are taken according to the market trend reversals. The market trend reversals indicate the potential of a market to experience an up-trend or downtrend, with the associated magnitude of the trend reversals. The direction of the trends is indicated by the sign of the function of market trend reversals potential, which can be derived by the mean of exponential moving average (EMA) and Moving Average Convergence Divergence (MACD) of share prices [46].

Price signal $y(t)$ is measured on the trading day t. The EMA imparts a higher weightage ω on recent prices near the current trading day t_c, that is calculated in Eq. (3):

$$\mathrm{EMA}_{t_c} = \omega(y(t_c)) + (1-\omega)\mathrm{EMA}_{t_c-1} \tag{3}$$

The value of weightage ω is derived in Eq. (4).

$$\omega = \frac{2}{k+1} \tag{4}$$

where k denotes the number of past trading days from the current trading day. Generally, the EMA is capable of providing a responsive indication of price trends and fluctuations.

Trend reversals are detected using the crossovers (i.e. intersections) of the MACD signal line [46, 47]. A MACD line is derived from the long-term EMA (26 periods) and the short-term EMA (12 periods) as illustrated in Eq. (5).

$$\mathrm{MACD} = \mathrm{EMA}[12] - \mathrm{EMA}[26] \tag{5}$$

The signal line is derived from the EMA[9], i.e. the 9 periods EMA. The crossovers are monitored to obtain insights of the price trends. During a buy crossover, where the MACD line intersects upwards with the signal line, it indicates that the period will be undergoing a bullish period. Conversely, during a sell crossover, where the MACD line intersects downwards with the signal line, it indicates that the period will be undergoing a bearish one. An advantage of the MACD is that both the momentum and trend can be determined in a single indicator.

2.2 State

A RL agent is able to interact with its environment at each time step t. The environment is represented in the form of a state s_t ε S, where S is the set of all available states [28]. The state vector at time t that is observed by the RL agent is updated in Eq. (6):

$$s_t = [\mathrm{EMA}_1, \mathrm{MACD}_1, \mathrm{EMA}_2, \mathrm{MACD}_2, \Delta t] \tag{6}$$

where EMA_1 is the standardised 6-day EMA of 15 days of the high risk index; EMA_2 is the standardised 6-day EMA of 15 days of the medium risk index; MACD_1 is the normalised 6-day difference in MACD line and signal line of the high risk index; MACD_2 is the normalised 6-day difference in MACD line and signal line of the medium risk index; and Δt is the difference in the number of days from the previous market trend reversal.

2.3 Action

With the current state s_t as input, the RL agent takes an action a_t ε $A(s_t)$, where $A(s_t)$ is the set of possible actions being taken in the state s_t. For each action, a reward r_t is received to evaluate the action outcomes, while the state will be moving into the state s_{t+1}. The modifications of the reward structure of the RL agent are limited to four different actions as follows:

(1) Increase high risk assets portfolio composition rate, and at the same time, reduce the composition of other assets.
(2) Increase medium risk assets portfolio composition rate, while reducing the composition of other assets.
(3) Increase both high and medium risk assets portfolio composition rate, while reducing the composition of low risk assets.
(4) Increase low risk assets portfolio composition rate, while reducing the composition of other assets.

2.4 Reward structure

In the environment, rewards are defined per step of the training. Each step refers to each market trend reversal detected in the period under study. The NAV reward comprises of two different components: the NAV change reward, and the current NAV reward.

For the first component, the NAV change reward is computed using Eq. (7):

$$\mathrm{Reward} = \frac{\mathrm{Changed\,composition} - \mathrm{Original\,composition}}{\mathrm{Original\,composition}} \times \mathrm{TSF} \tag{7}$$

where the Changed composition refers to the NAV obtained after a period of 10 days using the changed portfolio composition according to the actions of the RL agent; the Original composition is the NAV obtained after a period of 10 days using the original portfolio composition; and TSF denotes the Time Scaling Factor. The process is as shown in Fig. 1, where the difference between the Changed composition and the Original composition is visualised at time $t = 2$.

For the term of *TSF*, its value is determined using Eq. (8):

$$\mathrm{TSF} = 0.5 \times \frac{\text{Number of days past}}{\text{Total number of days}} + 0.5 \tag{8}$$

The TSF in Eq. (4) scales the NAV change rewards from 0.5 to 1.0, to reduce the degree of rewards achievable over time. This is to place more emphasis on initial actions as they have a greater impact on the final NAV due to the compounding effect.

For the second component, the current NAV reward is computed by simple division of the current NAV over a constant of 10,000,000 to normalise the reward, as shown in Eq. (9).

$$\text{current NAV reward} = \frac{\text{current NAV}}{1 \times 10^7} \tag{9}$$

where the current NAV is the NAV obtained after a period of 10 days using the changed portfolio composition. It is to normalise the current NAV reward. As such, the total reward per step is obtained as in Eq. (10).

$$\begin{aligned}\text{Total Reward} = {} & \text{NAV change reward} \\ & + \text{current NAV reward}\end{aligned} \tag{10}$$

Therefore, by referring to the rewards received per step, the RL agent will gain an insight to the performance of its immediate action based on the current state of the environment.

2.5 RL agent

For the RL Agent, a Q network is set up to determine the Q values of actions for each state [41, 45]. The Q network is shown in Fig. 2, where it consists of one input layer, one hidden layer and one output layer. The size of the hidden layer is 100 neurons.

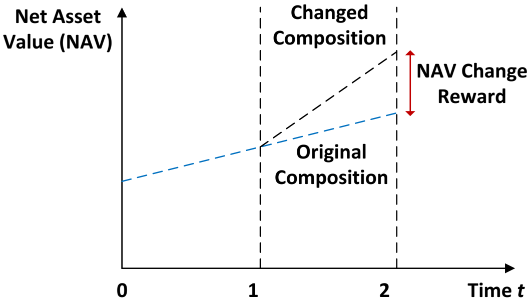


Fig. 1 Difference in NAV after composition change

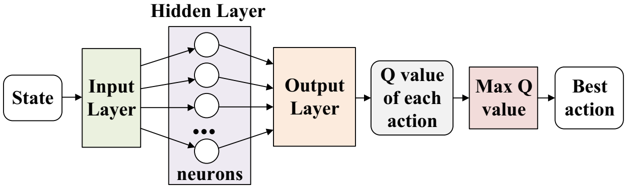


Fig. 2 Q network

The Q value of an action to be predicted by the Q network is determined by the Bellman equation [45], as shown in Eq. (11).

$$Q(s,a) = r(s,a) + \gamma \max_{a \in A(s)} \sum_{s'} P(s'|s,a) Q(s',a') \tag{11}$$

where s' is next state in the set of S states; a' is the next action; $Q(s', a')$ is the Q value for the next state s'; $r(s, a)$ is the rewards of the current state; the discount rate γ is the discount of the next Q value which is set at 0.99 in this paper; and $P(s' |s, a)$ is the probability of the state s' happening, given s and a which is set to a value of 1. Since the optimal policy is explored and followed by the agent, it aims to determine the best possible next action a' in the state s' to maximize the Q value [45], as shown in Fig. 2.

The portfolio rebalancing is performed according to the actions derived by the Q network of the agent. The NAV reward and change reward are measured according to the price changes of all assets in the portfolio after the rebalancing. The market trend reversals are monitored and computed using the updated values of the crossovers from the MCAD signals. It results in the triggering of next iterations of learning of the trading agent, as shown in the flowchart of Fig. 3.

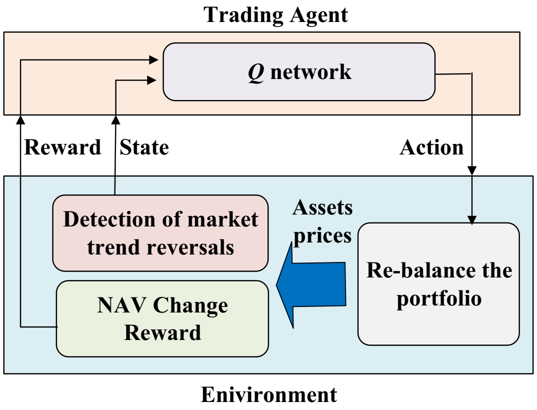


Fig. 3 Flowchart for Q learning updates

3 Proposed RL for dynamic rebalancing

The proposed RL agent is used to dynamically rebalance a portfolio. In order to study and evaluate the performance of the RL agent, four different combinations for portfolio switches and price prediction models are considered as follows.

Firstly, we will present the proposed RL agent with a full portfolio rebalancing, without price prediction. The full portfolio rebalancing method refers to the full change of the composition of asset reallocation to the highest setting to be guided by the highest potential. It aims to achieve a better return within a single trading day. This is done whenever trend reversals are detected.

Secondly, the proposed RL agent with a full portfolio rebalancing can be improved using the asset price prediction achieved by LSTM models. It is able to fine tune the technical indicator by centring and swifting the market information lag using the LSTM price prediction models.

Thirdly, the proposed RL agent is revised to incorporate a gradual portfolio rebalancing method without price predictions. The gradual portfolio rebalancing method changes the asset composition rates by $k\%$ per trading day, instead of complete change to highest composition rate for an asset in the single day.

Fourthly, the RL agent with gradual portfolio rebalancing method is further improved using the LSTM price prediction models. These four combinations of methods for the proposed RL agent will be described with the corresponding experiments' result analyses.

3.1 Experiment set-up

In order to better illustrate the performance analysis in the experiment, the proposed RL agent is used to dynamically rebalance three portfolios with good portfolio diversifications. One portfolio consists of global indices with assets at different risk levels. The other two portfolios comprise stock assets from different sectors in the U.S. market. These three portfolios are depicted as follows:

- The first portfolio includes three market index assets with varying degrees of market risk, namely the IBOVESPA Index (BVSP) which is a Brazillian stock index, the TSEC weighted index (TWII) which is a Taiwanese stock index, and the NASDAQ Composite (IXIC). This global selection of market indices in the portfolio encompasses three different indices with different market maturity. It provides different diversity of risks and financial market volatilities. The BVSP, TWII and IXIC indices are classified as high, medium, and low risk, respectively, considering the political environment and market volatility. The period 2014–2018 is chosen due to the presence of significant bullish and bearish market trends.
- The second portfolio consists of American Express Company (AXP), McDonald's Corp. (MCD), and Walmart Inc. (WMT) reported in [32] in the period from January 2016 to December 2018, which are selected from different sectors of S&P 500 (Standard and Poor's 500) and uncorrelated as much as possible. These three stocks are large-cap companies with market capitals at hundreds billion dollars.
- The third portfolio consists of UMB Financial Corp (UMBF), Uniti Group Inc. (UNIT), and Mandiant Inc. (MNDT, known as FEYE previously) from the NASDAQ market. They are selected from different sectors with different risk levels and different market trends mixing with bullish, bearish, and stagnant during the same time frame January 2016–December 2018. These three stocks have market capitals worth about 3–4 billion dollars.

The time frames are selected due the mixture of different market trends of the assets in the portfolios. It could enable more obvious observations on the process of dynamic portfolio rebalancing by the proposed algorithm, such that the performances can be better evaluated accordingly in the experiments.

Trend reversal periods are generated using the MACD, for which a buy crossover indicates an upwards trend reversal while a sell crossover indicates a downwards trend reversal. The experiments are set up with an initial NAV of $300,000 for each portfolio. It is initially split evenly by allocating $100,000 to each asset in each portfolio. The rebalancing approach is achieved through the combination of SAA and TAA strategies, with the base composition rates (BCR) given to each asset in the portfolio to prevent only keeping a single asset. It is to preserve the asset diversification of each portfolio. In the experiments, the BCR is set as 0.1 for each asset in the portfolio. A 0.125% commission rate is imposed on each transaction made during the executions of portfolio rebalancing. Tensorflow is used in the following RL experiments. The SciKit-Learn library is used for feature scaling in data preprocessing [48].

3.2 Proposed RL with full portfolio rebalancing method

The flowchart of the RL agent for full portfolio rebalancing method is shown in Fig. 4. The adjustments of the composition rates of the portfolio are completed in a single day once the market trend reversal is detected.

As indicated in the flowchart, the action policy of the RL agent for the full portfolio rebalancing is set as follows:

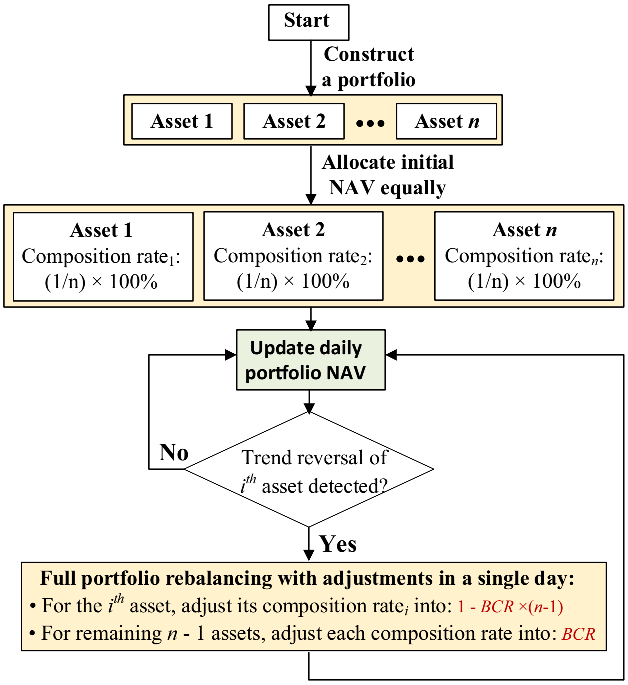

Start
Construct a portfolio
Asset 1
Asset 2
•••
Asset *n*
Allocate initial NAV equally
Asset 1 Composition rate1: (1/n) × 100%
Asset 2 Composition rate2: (1/n) × 100%
•••
Asset *n* Composition rate*n*: (1/n) × 100%
Update daily portfolio NAV
No
Trend reversal of *i*th asset detected?
Yes
Full portfolio rebalancing with adjustments in a single day:
• For the *i*th asset, adjust its composition rate*i* into: 1 - *BCR* ×(*n*-1)
• For remaining *n* - 1 assets, adjust each composition rate into: *BCR*

Fig. 4 Flowchart for full portfolio rebalancing

- Construct a portfolio consisting of a number of n assets according to the initial NAV being equally allocated to each asset.
- Upon the detection of an upward trend reversal for the ith asset ($1 \leq i < n$), increase the composition rate of this asset as shown in Eq. (12) in a single trading day,

$$\text{Composition rate}_i = 1 - \text{BCR} \times (n - 1) \tag{12}$$

where BCR refers to base composition rate as mentioned in Sect. 3.1. In the experiment, the BCR is set as 0.1. Hence, full allocation is set to the i-th asset.

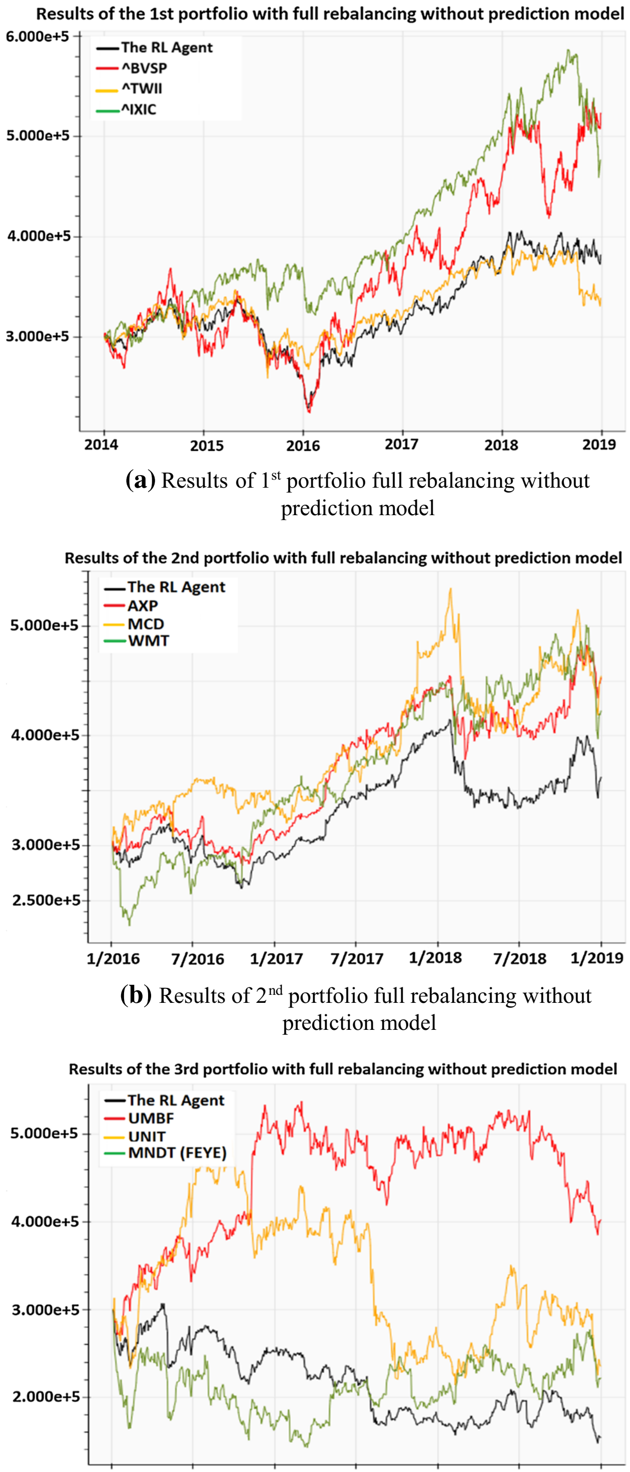
- Reduce the composition rate of each of the remaining assets to be at the BCR. This sets the allocation to the remaining assets at the BCR allocation setting.

In the experiments to dynamically rebalance portfolios of this paper, there are three assets in each portfolio, i.e. n = 3. As such, when an upward trend reversal of one asset is detected, the action policy is set to switch the composition rate of the corresponding asset to be 0.8, as derived in Eq. (13).

$$1 - \text{BCR} \times (n - 1) = 1 - 0.1 \times (3 - 1) = 0.8 \tag{13}$$

While the composition rates of the remaining two assets of the portfolios are reduced to 0.1 each, respectively.

The experiments have been performed for the RL agent with full portfolio rebalancing and no predictive models of the three portfolios. For the first portfolio consisting of three market indices assets, the experimental results are



(a) Results of 1st portfolio full rebalancing without prediction model

(b) Results of 2nd portfolio full rebalancing without prediction model

(c) Results of 3rd portfolio full rebalancing without prediction model

Fig. 5 **a** Results of 1st portfolio full rebalancing without prediction model. **b** Results of 2nd portfolio full rebalancing without prediction model. **c** Results of 3rd portfolio full rebalancing without prediction model shown in Fig. 5a, with the black plots representing the results of the portfolio NAV rebalanced by the RL agent. It is observed that the performance of the RL agent is not good, with a 20–25% difference in performance lower than those of the BVSP and IXIC indices. The RL agent fails to exploit the bullish trends present in the BVSP and IXIC indices. It fails to rebalance the portfolio to increase the composition rates of the BVSP and IXIC indices.

For the second portfolio consisting of three stock assets from S&P 500 reported in [32], the experiment results of full portfolio rebalancing without predictive modelling are shown in Fig. 5b. It is observed that the results of the RL agent in the black plots are not comparable to those of individual stock assets in this portfolio. The NAV performances of the RL agent are about 17–28% lower than those of WMT, AXP, and MCD in the second portfolio.

For the third portfolio comprising three stock assets from the NASDAQ market, the experiment results of full portfolio rebalancing are shown in Fig. 5c. The NAV results of the RL agent in the black plots suffer loss due to the very different market trends of each stock asset. The performance of the RL agent in full portfolio rebalancing is worse than that of individual stock assets in this portfolio, with about 16–116% lower than those of MNDT, UNIT and UMBF in the portfolio.

The experiment results are not satisfactory. The model of the proposed RL agent for full portfolio rebalancing needs be further examined and improved.

3.3 Market information lag

After inspections, one of the reasons to cause unacceptable performance in the RL model could be the information lag in the indicators used to detect market trend reversals. As shown in Fig. 6, the price trend prediction is performed based on the EMA of the historic prices in the past seven trading days, i.e. $0 \leq t \leq 6$. The notation $C1$ models the EMA window used in the experiment. The true centre of the computed EMA is actually at $t = 3$ instead of $t = 6$, at the day of action. Therefore, the derived price trend of the market at $t = 6$ in $C1$ is actually the true price trend of the market at $t = 3$, which is a 3-day lag in information. As such, for the indicator giving a better and more accurate prediction of the true market trends at $t = 6$, it is required to predict the prices for $t = 7$, $t = 8$ and $t = 9$. Therefore, it can judiciously compute the true EMA of stock prices at $t = 6$ as shown in $C4$ of Fig. 6.

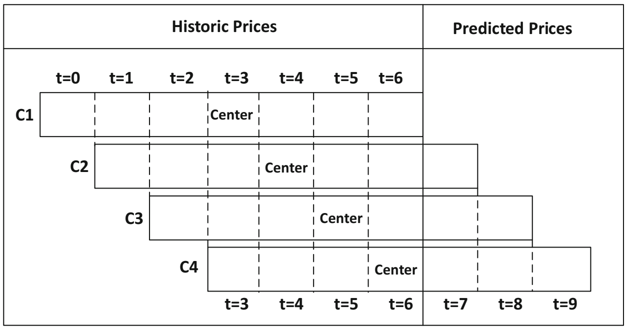


Fig. 6 Time lag of trend indicators

3.4 LSTM price prediction model

In order to verify the hypothesis of the RL agent having decreased performance due to the information lag in the market, a further study is conducted. In our research, LSTM models are trained to predict stock prices of the next three trading days, i.e. $t = 7$, $t = 8$ and $t = 9$. The prediction is conducted in order to reduce the time lag of the indicators of EMA and MACD used in the state of the RL environment. It can improve the information provided to the RL agent for its actions.

The LSTM model for price prediction is shown in Fig. 7. In this model, 32 LSTM units are used between the input and output layers, with a lookback period of seven days and a learning rate of 0.001. The loss is computed using the mean squared error approach. All features are normalized before being passed to the input layer.

In the experiments of the full portfolio rebalancing for three portfolios, the LSTM models are trained using their historic prices in the range of year 2014–2018. The prices of the next three days are predicted for each asset in the portfolios during the process of portfolio rebalancing. Using this method, the information lag is reduced, to improve the performance of the portfolio rebalancing.

3.5 Full portfolio rebalancing with LSTM price prediction

The set of experiments have been performed again for the three portfolios using the full portfolio rebalancing method with the LSTM prediction models by removing the 3-day market information lag.

The experiment results with the LSTM price prediction model for the first portfolio are shown in Fig. 8a. It is observed that the performance of the revised RL agent for

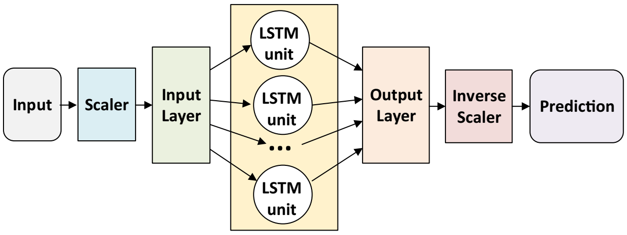
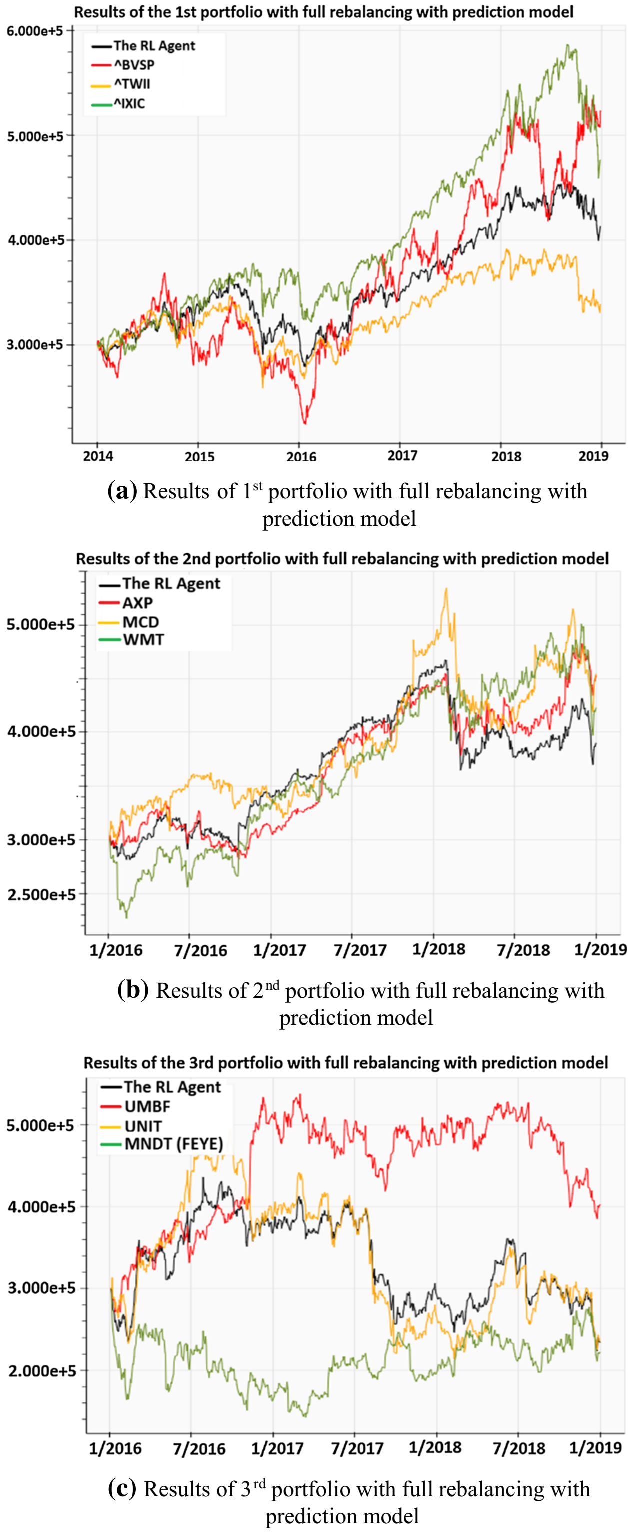


Fig. 7 LSTM model for stock price prediction



(a) Results of 1st portfolio with full rebalancing with prediction model

(b) Results of 2nd portfolio with full rebalancing with prediction model

(c) Results of 3rd portfolio with full rebalancing with prediction model

Fig. 8 **a** Results of 1st portfolio with full rebalancing with prediction model. **b** Results of 2nd portfolio with full rebalancing with prediction model. **c** Results of 3rd portfolio with full rebalancing with prediction model

full rebalancing with prediction model improved by about 5% compared to that of full rebalancing without prediction model. The RL agent is able to better rebalance the portfolio according to market trends, e.g. in the period of the beginning of 2015 and 2018. Therefore, the reduced time lag improves the performance of the full portfolio rebalancing. However, the performance of the RL agent is still below those of the BVSP and IXIC indices by about 15–20%.

For the second portfolio, the experiment results of full rebalancing with the price prediction model are shown in Fig. 8b. It is observed that the RL agent performs better than the method of full rebalancing without prediction model. The NAV results are better than individual stock assets in January 2017–October 2017. But its performance drops in year 2018 and becomes worse than those of AXP, MCD and WMT by about 11–21%.

For the experiments of the full rebalancing with prediction model for the third portfolio, the results are shown in Fig. 8c. It is observed that the performance of the RL agent is improved by about 11% compared to the method of full rebalancing without prediction model. Its NAV results are better than those of MNDT, but worse than those of UMBF by about 70%.

Further improvements to the proposed RL model are required to enhance the performance.

One aspect of the proposed RL agent to be looked into is the policy of actions. Currently, a full portfolio rebalancing method is used when a trend reversal is detected, where the composition rate of an asset can drastically step changed from the BCR to the value of $(1 - \mathrm{BCR} \times (n - 1))$ and vice versa. For example, if the portfolio consists of three assets, the composition rate of an asset changes drastically from 0.1 to 0.8 for the selected action. For the full portfolio rebalancing method, if the proposed RL agent chooses a wrong action, the penalty will be maximised in periods of uncertainty, due to the high commission fees, i.e. transaction costs, incurred in a big change in composition of assets.

3.6 Gradual portfolio rebalancing

Therefore, it would be better if the RL agent adopts a gradual change in composition rate instead of a full switch in composition from the BCR to the value of $(1 - \mathrm{BCR} \times (n - 1))$ within a single day. It will reduce the penalty of mistakes and improve results.

When a market trend reversal in the portfolio is detected, the revised RL actions for the gradual portfolio rebalancing method are as follows:

(1) Increase high risk asset portfolio composition rate by $k\%$ per day, where $0\% < k < (1 - \mathrm{BCR} \times (n - 1))$;

and reduce the composition of other assets, until reaching BCR or another trend reversal detected.

(2) Increase medium risk assets portfolio composition rate by $k\%$ per day; and reduce the composition of other assets, until reaching BCR or another trend reversal detected.

(3) Increase both high and medium risk assets portfolio composition rate by $0.5 \times k\%$ per day; and reduce the composition of low risk assets, until reaching BCR or until another trend reversal detected.

(4) Increase low risk assets portfolio composition rate by $k\%$ per day; and reduce the composition of other assets, until reaching BCR or until another trend reversal detected.

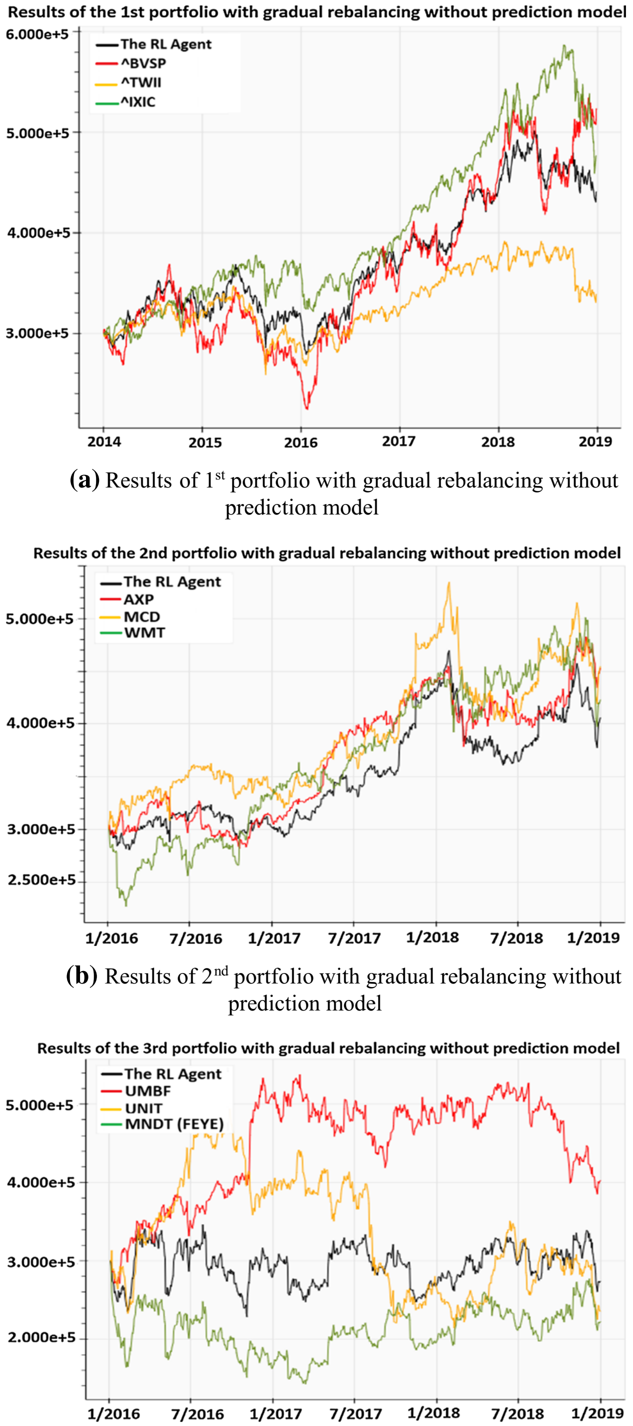
Using the gradual portfolio rebalancing method, when a market trend reversal is detected, the changes of the composition rate for the corresponding asset are set by $k\%$ per day. The portfolio rebalancing will continue for few days, until this asset reaches the value of $(1 - \text{BCR} \times (n - 1))$ or the remaining assets reach the BCR, or another trend reversal is detected.

As such, changes in the portfolio compositions are less sudden. Continuous trend reversals within short time intervals will have a lesser penalty in commission charges due to a slower composition change. Additionally, mistakes made by the RL agent will be less costly and less commission is incurred as well, if the next trend reversal is close by.

However, using this gradual portfolio rebalancing method, the reactions to significant short term bullish trends and bearish trends will be slower. The proposed RL agent may not be able to rebalance the portfolio fast enough to either exploit the large bullish trend or protect against the large bearish trends. Therefore, it is a trade-off between the value of $k\%$ changes per day and the reaction latency. After a few experiments for fine tuning of the value of k with three assets in the portfolio, the gradual portfolio rebalancing method exhibits better performance when $k\%$ is set at 30%. As such, in the experiment of the gradual portfolio rebalancing method of this paper, the changes of the composition rate for the corresponding asset are set as 30% per day.

Two more sets of experiments have been conducted for the gradual portfolio rebalancing method; one is without the LSTM price prediction model. The other is with the price prediction model to remove the 3-day market information lag.

The experiments of the RL agent for the gradual portfolio rebalancing method without prediction model for the three portfolios are conducted, which are discussed in the next three paragraphs.



(a) Results of 1st portfolio with gradual rebalancing without prediction model

(b) Results of 2nd portfolio with gradual rebalancing without prediction model

(c) Results of 3rd portfolio with gradual rebalancing without prediction model

Fig. 9 **a** Results of 1st portfolio with gradual rebalancing without prediction model. **b** Results of 2nd portfolio with gradual rebalancing without prediction model. **c** Results of 3rd portfolio with gradual rebalancing without prediction model

For the first portfolio, Fig. 9a shows that the RL agent is able to adopt a more risk adverse approach by increasing the composition rate of the IXIC index asset at the end of

2014, which prevents further decreases in the portfolio NAV. Additionally, at the beginning of 2018, it adopts a risk seeking stance by increasing the portfolio composition of the BVSP index asset, which led to a substantial increase in NAV. However, at the end of 2018, it does not rebalance to the IXIC index asset in time to exploit the rapid rise in the IXIC index. Overall, the RL agent has an improved performance and a better risk profile, as its movements are less volatile than that of the BVSP index of the first portfolio.

The experiment results of the second portfolio are shown in Fig. 9b. The NAV results of the RL agent in black plots for gradual portfolio rebalancing without the prediction model are better than those of the full portfolio rebalancing without prediction model by about 12%; while they are marginally better than those of the full portfolio rebalancing with prediction model.

For the experiments of the third portfolio, the results are shown in Fig. 9c. It is observed that the RL agent performs better than its results for full portfolio rebalancing without prediction model by about 41%; and marginally better than those of the full portfolio rebalancing with prediction model.

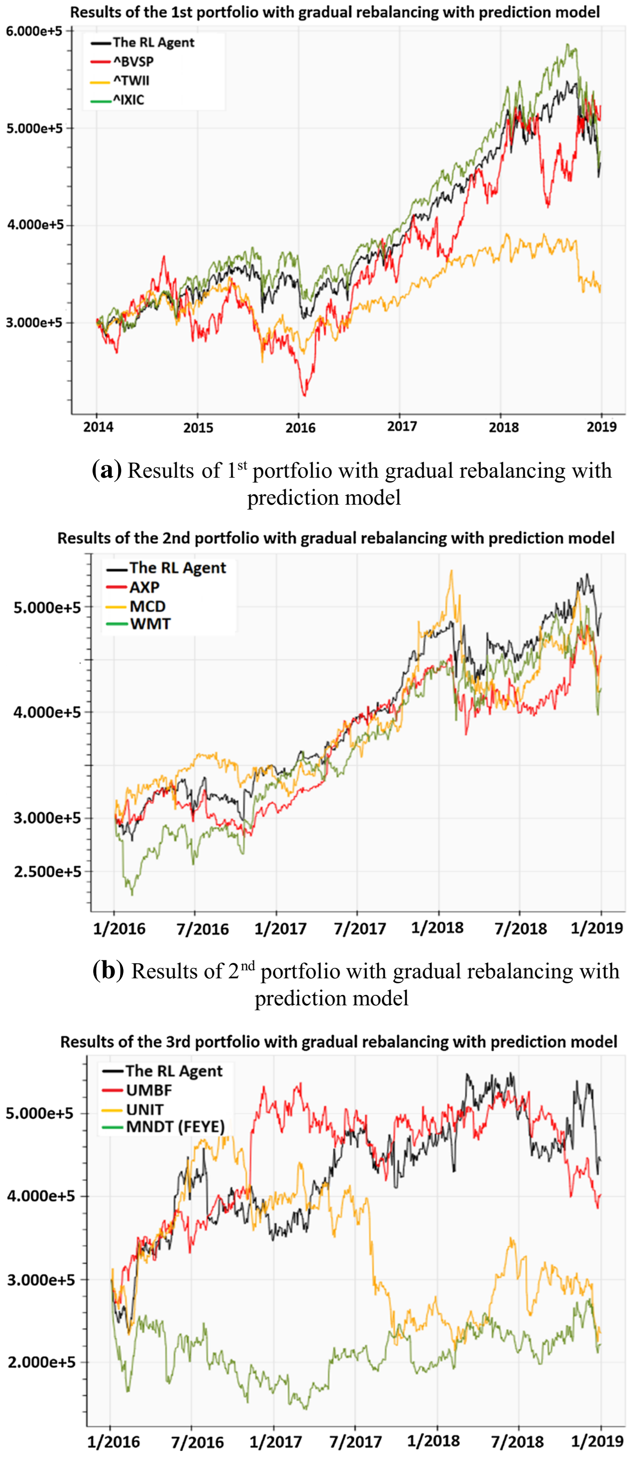
Similarly, for the gradual portfolio rebalancing method with the prediction model, the experiments are performed for the three portfolios, presented in the next three paragraphs.

The experiment results of the first portfolio using the gradual portfolio rebalancing with the LSTM prediction model are shown in Fig. 10a. It is observed that the proposed RL agent using the prediction model further improves the performance of the portfolio NAV by around 5%, better than that achieved without the LSTM prediction model. Figure 10a shows a trend for the RL agent to increase the portfolio composition of the IXIC index asset, which leads to the similarity in the shape of the portfolio NAV curve and that of the IXIC index.

For the second portfolio managed by the RL agent using gradual portfolio rebalancing with the prediction model, the experiment results are shown in Fig. 10b. It is observed that the RL agent performs better than all individual stock assets in this portfolio by about 8–14%. The RL agent rebalances the composition rates of the portfolio successfully to ride the upwards market trends in year 2017 and after February 2018.

For experiments of the third portfolio, the results of the RL agent using gradual portfolio rebalancing with the prediction model are shown in Fig. 10c. The RL agent outperforms each individual stock assets in year 2018 by about 9–100%.

Observed from Fig. 10a–c, the RL agent using gradual portfolio rebalancing with the prediction model outperforms the full portfolio rebalancing method in the previous experiments by about 27.9–93.4%. It means that the effect of the reduction in the portfolio NAV by the commission



(a) Results of 1st portfolio with gradual rebalancing with prediction model

(b) Results of 2nd portfolio with gradual rebalancing with prediction model

(c) Results of 3rd portfolio with gradual rebalancing with prediction model

Fig. 10 **a** Results 1st portfolio with gradual rebalancing with prediction model. **b** Results 2nd portfolio with gradual rebalancing with prediction model. **c** Results 3rd portfolio with gradual rebalancing with prediction model fees is larger than the effect of the RL agent not being able to exploit bullish trends and protect against bearish trends.

3.7 Discussions

As discussed previously in Sect. 3, there are four different methods of the proposed RL agent, i.e. (1) full portfolio balancing without LSTM prediction model, (2) full portfolio balancing with LSTM prediction model, (3) gradual portfolio balancing without LSTM prediction model, and (4) gradual portfolio balancing with LSTM prediction model. Experiments have been performed using the three portfolios consisting of market index assets and stock assets with different risk levels from different sectors. It is observed from the experiment results that the RL rebalanced portfolios are able to switch among assets according to the market trends of each asset, to increase the profits, while considering the corresponding market risks in the experiment period.

In order to better visualise the quantitative performance enhances, the comparison results of these four RL methods for the first portfolio are shown in Table 1. The comparison results for the second portfolio and the third portfolio are shown in Tables 2 and 3, respectively.

The percentage of NAV return at the end of 2018 is referred as the increment of the portfolio NAV at the end of 2018 based on the initial NAV at the beginning of the experiment period. It is computed in Eq. (14), where the initial NAV is \$300,000.

$$\text{NAV return } \% = \frac{\text{NAV at the end of 2018} - \text{initial NAV}}{\text{initial NAV}} \tag{14}$$

The other variable shown in Tables 1, 2 and 3 is the percentage of NAV max drop in the experiment period, which is referred as the percentage of maximum NAV decrement in the time frame. Its calculation is shown in Eq. (15).

$$\text{NAV max drop } \% = \frac{\text{Lowest NAV in the period} - \text{initial NAV}}{\text{initial NAV}} \tag{15}$$

It is observed from Tables 1, 2 and 3 that the performances of NAV portfolio managed by the four different types of RL agent keep improving with better NAV return and enhanced maximum decrement in the experiment period.

Observed in Table 1 for the 1st portfolio, although their percentages of the NAV return are lower than that of the BVSP index, the methods of the RL agent show better percentage of NAV max drop and much lower volatility. Here, we compare the RL rebalancing strategies against simple Buy and Hold strategies [12] of the three underlying assets; namely BVSP, TWII and IXIC. It illustrates the capability of the proposed RL agent to maximum the profits and handle well high risk assets, such as the BVSP index asset.

It is shown in Table 2 for the 2nd portfolio, the four RL agents exhibit the same trends to improve the NAV return of the portfolio. The gradual portfolio rebalancing with the LSTM prediction model achieves the best returns at 63.3% than individual assets of AXP, MCD, WMT in this portfolio, as well as better than those of AXP, MCD, and WMT reported in [32] in considering their trading hourly returns and corresponding portfolio weights.

The 3rd portfolio results shown in Table 3 indicate the same trend of the NAV returns for the four RL agents, with

Table 1 Difference in NAV return and max drop for the 1st portfolio

	(1) Full rebalance without prediction (%)	(2) Full rebalance with prediction (%)	(3) Gradual rebalance without prediction (%)	(4) Gradual rebalance with prediction (%)
NAV Return % by end of 2018				
RL agent	27.3	39.7	47.2	55.2
BVSP	74.3			
TWII	12.7			
IXIC	48.7			
NAV max drop %				
RL agent	−23.7	−6.8	−6.5	0.7
BVSP	−25.3			
TWII	−14.3			
IXIC	−4.7			

Table 2 Difference in NAV return and max drop for the 2nd portfolio

	(1) Full rebalance without prediction (%)	(2) Full rebalance with prediction (%)	(3) Gradual rebalance without prediction (%)	(4) Gradual rebalance with prediction (%)
NAV Return % by end of 2018				
RL agent	21.7	29.3	36.7	63.3
AXP	51.6			
MCD	51.7			
WMT	40.7			
NAV max drop %				
RL agent	−14.6	−6.7	−6.6	−6.9
AXP	−6.3			
MCD	−1.7			
WMT	−17.5			

Table 3 Difference in NAV return and max drop for the 3rd portfolio

	(1) Full rebalance without prediction (%)	(2) Full rebalance with prediction (%)	(3) Gradual rebalance without prediction (%)	(4) Gradual rebalance with prediction (%)
NAV Return % by end of 2018				
RL agent	−46.7	−20.7	−8.7	46.7
UMBF	34.3			
UNIT	−20.6			
MNDT	−25.7			
NAV max drop %				
RL agent	−49.3	−25.3	−22.7	−20.5
UMBF	−9.7			
UNIT	−26.7			
MNDT	−52.3			

the gradual portfolio rebalancing with the LSTM prediction model obtaining the best returns. It is also observed that this portfolio exhibits larger volatility with negative returns and percentages of NAV drops of the portfolio assets. It is caused by the mixture of different market trends of the individual stock asset in the experiment time frame.

The RL-based approach to balance a portfolio consisting of different risk assets allows the opportunistic attempts to benefit from market trend reversals. The main disadvantage of such an opportunistic strategy is the commission leak which reduced the overall NAV. This will fail if any of underlying assets experience prolonged bullish market trend where a simple Buy and Hold strategy works better over that period, since there are lesser commission leaks. The gradual portfolio rebalancing with the prediction model suffers the least maximum NAV drop when compared against the simple Buy and Hold strategy for the three constituent assets, as well as the other three variants of portfolio rebalancing strategies as shown in Tables 1, 2 and 3.

4 Conclusions

In this paper, the proposed RL agent has demonstrated the ability to dynamically adjust the portfolio composition rates according to the market trends, risks and returns of each asset throughout the periods under study. Four different methods for the proposed RL agent are discussed and evaluated including full and gradual portfolio rebalancing, without and with price prediction models on technical indicator centring. The experiments for these four methods have been conducted and evaluated using three portfolios: one portfolio for market index assets, the other two portfolios are stock assets of NYSE and NASDAQ market. Observed from the experiment results presented in Figs. 5, 8, 9, 10 and Tables 1, 2 and 3, the RL agent with gradual portfolio rebalancing with LSTM prediction model performs better, as it uses the appropriate trading behaviours in gradually adjusting the portfolio composition, instead of switching portfolio compositions in a single trading day. The performance improvements of the gradual rebalancing with prediction model are achieved at about 27.9–93.4% over the full rebalancing without prediction model. Its performances are also higher than most of the individual assets in these three portfolios, except for the BVSP market index.

The experiments illustrate that a properly tuned RL agent with and without a LSTM price prediction model to centre technical indicators can utilise dynamic rebalancing with adjusting risks to improve portfolio returns. Thus, the strategy of dynamic portfolio rebalancing with vigorous risks coupled with the concepts of SAA and TAA strategies is shown to work well by the proposed algorithm.

Future works regarding the RL agent can try to improve the stock prediction model that is used to reduce the time lag for technical indicators. It aims to examine the issue of optimising the number of trades to reduce commission leak and examine the effectiveness of other techniques of RL such as the Actor-Critic, Experience Replay or Double Q-learning in dynamic portfolio rebalancing.

Funding Not applicable.

Availability of data and material Data are available upon request.

Code availability Code is available upon request.

Declarations

Conflict of interest The authors declare no conflict of interest.

Consent for publication Yes. All authors agree for the publication.

References

1. Almahdi S, Yang S (2017) An adaptive portfolio trading system: a risk-return portfolio optimization using recurrent reinforcement learning with expected maximum drawdown. Expert Syst Appl 87:267–279. https://doi.org/10.1016/j.eswa.2017.06.023
2. Lekovic M (2018) Investment diversification as a strategy for reducing investment risk. Ekonomski Horizonti 20:169–184. https://doi.org/10.5937/ekonhor1802173L
3. Chow T-M, Hsu JC, Kuo L-L, Li F (2014) A study of low-volatility portfolio construction methods. J Portf Manag 40(4):89–105
4. Statman M (2014) Behavioral finance: finance with normal people. Borsa Istanbul Rev 14(2):65–73
5. Faber MT (2007) A quantitative approach to tactical asset allocation. J Wealth Manag 9(4):69–79
6. Anson MJ (2004) Strategic versus tactical asset allocation. J Portf Manag 30(2):8–22
7. Brown R (2019) TAA properly defined. Finance Mark. https://doi.org/10.18686/fm.v0.1097
8. Tuyon J, Ahmad Z (2016) Behavioural finance perspectives on malaysian stock market efficiency. Borsa Istanbul Rev 16(1):43–61
9. Tversky A, Kahneman D (1992) Advances in prospect theory: cumulative representation of uncertainty. J Risk Uncertain 5(4):297–323
10. Liu Y, Nacher J, Ochiai T, Martino M, Altshuler Y (2014) Prospect theory for online financial trading. PLoS ONE. https://doi.org/10.1371/journal.pone.0109458
11. Yang L (2019) Loss aversion in financial markets. J Mech Inst Des 4(1):119–137
12. Cheong D, Kim Y, Byun H, Oh K, Kim T (2017) Using genetic algorithm to support clustering-based portfolio optimization by investor information. Appl Soft Comput. https://doi.org/10.1016/j.asoc.2017.08.042
13. Mashrur A, Luo W, Zaidi NA, Robles-Kelly A (2020) Machine learning for financial risk management: a survey. IEEE Access 8:203203–203223. https://doi.org/10.1109/ACCESS.2020.3036322
14. Matsumura K, Kakinoki H (2014) Portfolio strategy optimizing model for risk management utilizing evolutionary computation. Electron Commun Jpn 97(8):45–62
15. Lwin K, Qu R, Kendall G (2014) A learning-guided multi-objective evolutionary algorithm for constrained portfolio optimization. Appl Soft Comput 24:757–772. https://doi.org/10.1016/j.asoc.2014.08.026
16. Quintana D, Denysiuk R, Garcia-Rodriguez S, Gaspar-Cunha A (2017) Portfolio implementation risk management using evolutionary multiobjective optimization. Appl Sci 7(10):1079
17. Ahn W, Cheong D, Kim Y, Oh KJ (2018) Developing an enhanced portfolio trading system using k-means and genetic algorithms. Int J Ind Eng Theory Appl Pract 25(5)
18. Lim S, Kim M, Ahn C (2020) A genetic algorithm (GA) approach to the portfolio design based on market movements and asset valuations. IEEE Access. https://doi.org/10.1109/ACCESS.2020.3013097
19. Liu C, Yin Y (2018) Particle swarm optimised analysis of investment decision. Cogn Syst Res 52:685–690. https://doi.org/10.1016/j.cogsys.2018.07.032
20. Silva Y, Herthel A, Subramanian A (2019) A multi-objective evolutionary algorithm for a class of mean-variance portfolio selection problems. Expert Syst Appl 133:225–241
21. Liagkouras K, Metaxiotis K (2018) Multi-period meanvariance fuzzy portfolio optimization model with transaction costs. Eng

Appl Artif Intell 67:260–269. https://doi.org/10.1016/j.engappai.2017.10.010

22. Jiang Z, Liang J (2017) Cryptocurrency portfolio management with deep reinforcement learning. In: IEEE intelligent systems conference, pp 905–913. https://doi.org/10.1109/IntelliSys.2017.8324237
23. Du X, Zhai J, Lv K (2016) Algorithm trading using q-learning and recurrent reinforcement learning. Positions 1:1
24. Kristjanpoller W, Michell K (2018) A stock market risk forecasting model through integration of switching regime, ANFIS and GARCH techniques. Appl Soft Comput 67:106–116
25. Serrano W (2021) The random neural network in price predictions. Neural Comput Appl. https://doi.org/10.1007/s00521-021-05903-0
26. Sutton R, Barto A (2018) Reinforcement learning: an introduction. MIT Press, Cambridge
27. Garí Y, Monge D, Pacini E, Mateos C, Garino C (2021) Reinforcement learning-based application autoscaling in the cloud: a survey. Eng Appl Artif Intell. https://doi.org/10.1016/j.engappai.2021.104288
28. Baker B, Kanitscheider I, Markov T, Wu Y, Powell G, McGrew B, Mordatch I (2020) Emergent tool use from multi-agent autocurricula. In: International conference on learning representations
29. Krasheninnikova E, García J, Maestre R, Fernández F (2019) Reinforcement learning for pricing strategy optimization in the insurance industry. Eng Appl Artif Intell 80:8–19. https://doi.org/10.1016/j.engappai.2019.01.010
30. Sánchez E, Clempner J, Poznyak A (2015) A priori-knowledge/actor-critic reinforcement learning architecture for computing the mean–variance customer portfolio: the case of bank marketing campaigns. Eng Appl Artif Intell 46:82–92. https://doi.org/10.1016/j.engappai.2015.08.011
31. Soleymani F, Paquet E (2020) Financial portfolio optimization with online deep reinforcement learning and restricted stacked autoencoder: DeepBreath. Expert Syst Appl 156
32. Aboussalah A, Lee C (2020) Continuous control with stacked deep dynamic recurrent reinforcement learning for portfolio optimization. Expert Syst Appl 140
33. Huotari T, Savolainen J, Collan M (2020) Deep reinforcement learning agent for S&P 500 stock selection. Axioms. https://doi.org/10.3390/axioms9040130
34. Lee J, Kim R, Yi S, Kang J (2020) MAPS: multi-agent reinforcement learning-based portfolio management system. In: Twenty-ninth international joint conference on artificial intelligence, special track on AI in FinTech. https://doi.org/10.24963/ijcai.2020/623
35. Pendharkar PC, Cusatis P (2018) Trading financial indices with reinforcement learning agents. Expert Syst Appl 103:1–13
36. Hirchoua B, Ouhbi B, Frikh B (2021) Deep reinforcement learning based trading agents: risk curiosity driven learning for financial rules-based policy. Expert Syst Appl. https://doi.org/10.1016/j.eswa.2020.114553
37. Park H, Sim M, Choi D (2020) An intelligent financial portfolio trading strategy using deep Q-learning. Expert Syst Appl. https://doi.org/10.1016/j.eswa.2020.113573
38. Lei K, Zhang B, Li Y, Yang M, Shen Y (2020) Time-driven feature-aware jointly deep reinforcement learning for financial signal representation and algorithmic trading. Expert Syst Appl. https://doi.org/10.1016/j.eswa.2019.112872
39. Li Y, Zheng W, Zheng Z (2019) Deep robust reinforcement learning for practical algorithmic trading. IEEE Access. https://doi.org/10.1109/ACCESS.2019.2932789
40. Meng T, Khushi M (2019) Reinforcement learning in financial markets. Data. https://doi.org/10.3390/data4030110
41. Jeong G, Kim H (2019) Improving financial trading decisions using deep q-learning: predicting the number of shares, action strategies, and transfer learning. Expert Syst Appl 117:125–138
42. García-Galicia M, Carsteanu A, Clempner J (2019) Continuous-time reinforcement learning approach for portfolio management with time penalization. Expert Syst Appl 129:27–36. https://doi.org/10.1016/j.eswa.2019.03.055
43. Lucarelli G, Borrotti M (2020) A deep Q-learning portfolio management framework for the cryptocurrency market. Neural Comput Appl 32:17229–17244. https://doi.org/10.1007/s00521-020-05359-8
44. Serrano W (2020) Genetic and deep learning clusters based on neural networks for management decision structures. Neural Comput Appl 32:4187–4211. https://doi.org/10.1007/s00521-019-04231-8
45. Russell S, Norvig P (2010) Artificial Intelligence: a modern approach. Prentice Hall
46. DayTrading.com, “MACD: moving average convergence divergence,” [Online]: https://www.daytrading.com/macd
47. Tan J, Zhou W, Quek C (2015) Trading model: self reorganizing fuzzy associative machine: forecasted MACD-Histogram (SeroFAM-fMACDH). In: International joint conference on neural networks (IJCNN), pp 1–8. https://doi.org/10.1109/IJCNN.2015.7280571
48. Scikit-learn developers, “Preprocessing data”, updated in 2020. https://scikit-learn.org/stable/modules/preprocessing.html