Technological Forecasting & Social Change 198 (2024) 122944

Contents lists available at ScienceDirect

Technological Forecasting & Social Change

journal homepage: www.elsevier.com/locate/techfore

Multi-period portfolio optimization using a deep reinforcement learning hyper-heuristic approach

Tianxiang Cui [a,b,*], Nanjiang Du [a,b], Xiaoying Yang [a,b], Shusheng Ding [c]

[a] School of Computer Science, University of Nottingham Ningbo China, 199 Taikang East Road, Ningbo, 315100, Zhejiang, China
[b] Ningbo Key Laboratory of Digital Port Technologies, 199 Taikang East Road, Ningbo, 315100, Zhejiang, China
[c] School of Business, Ningbo University, 818 Fenghua Road, Ningbo, 315211, Zhejiang, China

ARTICLE INFO

Keywords:
Portfolio optimization
Deep reinforcement learning
Hyper-heuristic
Decision making
Uncertainty

ABSTRACT

Portfolio optimization concerns with periodically allocating the limited funds to invest in a variety of potential assets in order to satisfy investors' appetites for risk and return goals. Recently, Deep Reinforcement Learning (DRL) has shown its promising capabilities in sequential decision making problems. However, traditional DRL algorithms directly operate in the space of low-level actions, which exhibits poor scalability and becomes intractable in real-world problem instances when the dimensionality of the environment increases. To deal with this, in this work, a novel DRL hyper-heuristic framework is proposed for multi-period portfolio optimization problem. Instead of exploiting the entire action domain, our proposed approach is more effective by searching for low-level well-developed trading strategies. In addition, our proposed approach is data-driven and respects the nature of the problem by taking advantage of expert domain knowledge and posing it multidimensional states to further leverage additional diverse information from alternative views of the environment. The proposed approach is evaluated on five real-world capital market problem instances and numerous experimental results demonstrate our proposed method can achieve notable performance gains compared to state-of-art trading strategies as well as traditional DRL baseline method. The data we used are from five stock indices, covering the period from the 2012 to 2022. Our study can have salient policy implications for investment strategy formulation and effective regulatory frameworks establishment.

1. Introduction

Portfolio optimization plays a vital role in investment companies, hedge funds, banks and other financial institutions. Since the pioneering work of Markowitz Modern Portfolio Theory (MPT) framework (Markowitz, 1952), it has received sustained attention from both asset liability professionals and academics. Essentially, it can be considered the process of periodically reallocating the limited funds to invest in a variety of financial assets in order to satisfy investors' appetites for risk and return goals. With different practical trading constraints involved (Woodside-Oriakhi et al., 2011), the problem becomes a classical NP-hard Combinatorial Optimization Problem (COP) in operational research. Existing studies have largely focused on model-driven approaches. The general process is to use mathematical formulation to establish a scientific model first and then apply various optimization algorithms to solve the model. Typically, there exist two main families of approaches for solving model-based portfolio optimization problems. Exact algorithms, such as Branch-and-Bound (Bertsimas and Shioda, 2009) and Lagrangian relaxation framework (Shaw et al., 2008), are based on the clever and complete enumeration of the solution space. These algorithms can eventually obtain the optimal solution, but they may be prohibitive for solving large problem instances because of the exponential time complexity. Alternatively, approximation algorithms, such as metaheuristics (Woodside-Oriakhi et al., 2011) and hybridization methods (Cui et al., 2014), can generate good-quality solutions with small computational effort, but once the problem statement changes slightly, they need to be revised. In fact, one main issue of the model-based approach is that it normally focuses on the deterministic variants of the problem, in which some strong assumptions are often pre-setup in the model. For example, in classical MPT based portfolio optimization model, it often assumes the perfect information of the market can be obtained with perfect accuracy by financial analysts (Wu et al., 2022). Since there are so many potential sources that could effect the estimation, it is difficult to even get precise estimates for them in practice (Bonami and Lejeune, 2009). As a result, the algorithms developed using the model-driven approach may be hard to deploy in the unpredictable real-world financial markets because of the high

* Corresponding author.
E-mail address: tianxiang.cui@nottingham.edu.cn (T. Cui).

https://doi.org/10.1016/j.techfore.2023.122944
Received 13 March 2023; Received in revised form 7 September 2023; Accepted 20 October 2023
Available online 2 November 2023

level of uncertainty. Although some modeling techniques like stochastic programming can address this issue partially (Cui et al., 2020), they often lead to extremely large models that tend to be intractable for most practical problem instances.

With the advances in sequence-to-sequence learning (Sutskever et al., 2014) and the rise of computing power in recent decades, using Deep Learning (DL) methods for portfolio optimization problems has been revisited to exploit the potential laws of the market. The idea is to train a neural network for the price movements or trends prediction in a supervised manner using different techniques, such as textual analysis (Boubaker et al., 2021) and sentiment analysis (Eachempati et al., 2021). Then the solutions can be subsequently constructed from the problem specification. It has been shown that the trend forecasting approach is not guaranteed to obtain an optimal portfolio since the prediction loss is in different with the overall objective of the problem (Moody and Saffell, 2001). In addition, the performance of these methods heavily depends on the degree of prediction accuracy. Given the efficient market theory (Fama, 1965, 1970), there is no reliable way to predict the future path of stock price movement since all the relevant information shall be reflected in the stock price. The theoretical mechanism behind the efficient market theory centers on the concept of competition. In efficient markets, competition among investors ensures that prices quickly adjust to reflect new information (Laffont and Maskin, 1990). Rational investors actively participate in buying and selling securities, which drives prices towards their fundamental values. Any deviation from the fair price would create an opportunity for arbitrage, attracting investors who would quickly exploit the discrepancy and bring prices back in line with underlying values. This constant competition and price adjustment mechanism prevent the persistence of mispricing and, as a result, ensure market efficiency (Lamont and Thaler, 2003).

Nevertheless, tremendous studies have attempted to use financial variables to predict future stock prices, such as book-to-market ratios and dividend yields (Thaler, 1999). On this basis, Ang and Bekaert (2007) demonstrates the trivial predictive power of dividend yield in stock price prediction. In addition, Avramov (2002) points out that the predictability of stock prices is highly dependent on the model specification. More recently, Pyun (2019) reveals that the predictability of stock prices is time-varying, and some sample periods retain higher predictability than other sample periods. As a result, the accurate market trend prediction is extremely challenging due to the highly noisy, stochastic and chaotic nature of the financial market (Abedin et al., 2021; Efat et al., 2022; Shajalal et al., 2023), especially for portfolio constructions considering model specification and sample selection. Also, the non-stationary nature of market may further induce many sources of uncertainty (Ding et al., 2023), which will often cause a distribution shift between historical and future data. Moreover, the DL based models normally do not have interaction with the market, thus lack adaptivity to the real-world financial environment. In recent years, the efficient market theory has been applied into the cryptocurrency markets (Chu et al., 2019; Le Tran and Leirvik, 2020; Kang et al., 2022). Other scholars have discovered the time-varying characteristics for market efficiency in recent studies (Okoroafor and Leirvik, 2022).

According to the efficient market theory, the efficient market can correct the mispricing, which eliminates the arbitrage opportunities. It thereby motivates our study since our study aims to formulate the effective portfolios based on multiple trading strategies, which can yield higher returns. Instead of using trend prediction, our study proposes a deep reinforcement learning (DRL) hyper-heuristic approach to narrow down the search space and improve portfolio optimization. By incorporating domain knowledge and low-level trading strategies, the proposed approach goes beyond pure observations and aims to enhance asset allocation decisions in real-world financial markets. By using a DRL hyper-heuristic method, the study demonstrates how different trading strategies can be utilized in combination to achieve better performance. The empirical results of the study complement the efficient market theory by showcasing the trading opportunities that can be captured through portfolio construction. Our empirical results thereby can complement the efficient market theory to demonstrate the trading opportunities that can be grasped by the portfolio construction, which can be a channel that eliminate the market mispricing.

Recently, Reinforcement Learning (RL) algorithms have been proved effective in sequential decision making problems. The integration of DL and RL (DRL), is receiving a lot of attention due to its outstanding accomplishments across several fields (Silver et al., 2017; Vinyals et al., 2019; Jumper et al., 2021). Generally, RL can be viewed as an approximation of Dynamic Programming (DP) (Bellman, 1957) which is a general divide-and-conquer technique for solving complex problems by decomposing them into several different parts (sub-problems) that have a recursive relationship. After each part has been solved, DP provides a systematic procedure for determining the combination of results of the sub-problems in order to obtain an overall solution. Technically, a COP can be transferred to an equivalent DP problem. A DP problem has two main properties: overlapping subproblems and optimal substructure. The fundamental mathematical model in RL, Markov Decision Process (MDP), can satisfy both of the two properties. Consequently, RL can provide an appropriate paradigm for finding solutions for COPs (in the equivalent DP formulations). As a matter of fact, DRL has already shown promising abilities to obtain high-quality solutions to some COPs (Kong et al., 2019; Mazyavkina et al., 2021).

Building a successful portfolio within a vast problem space is difficult since there are hundreds of assets on a stock market. Existing DRL-based approaches directly operate in the space of low-level actions (i.e. asset weights), which can easily suffer from the "curse of dimensionality", becoming less efficient as the dimensionality of the environment increases. Motivated by the demand of solving large challenging real-world problems, various techniques that can narrow regions of the search space of DRL have been investigated, such as Monte Carlo Tree Search (MCTS)-based approach (Silver et al., 2016; Lee et al., 2018), rule-based approach (Radaideh and Shirvan, 2021), hyper-heuristic based approach (Zhang et al., 2021). Hyper-heuristics (Burke et al., 2019), which can be broadly referred to as "heuristics to select or generate heuristics", have been studied among them with the main goal of enhancing generality in performance across various problem instances and problem domains. Unlike meta-heuristic that operates directly on specific solutions, hyper-heuristic operates on the heuristic space, which makes it capable of handling cross-domain optimization (Pillay and Qu, 2018). A well-designed hyper-heuristic can often be applied to many different problem instances and problem domains with little or no modification, which can be particularly useful in complex or poorly understood environments like the financial market. Moreover, hyper-heuristics can provide robust solutions, as they are able to switch between different heuristics as required, making them less susceptible to the weaknesses of any single heuristic. The nature of the hyper-heuristic allows the DRL agent to adapt its trading strategy based on the current market situation. The hyper-heuristic framework has demonstrated great success in many traditional COPs (Rahimian et al., 2017; Ahmed et al., 2019) and consequently can serve as a promising framework for real-world complex online portfolio optimization problems.

In this work, a novel DRL hyper-heuristic framework is proposed to tackle real-world multi-period portfolio optimization problems. The following contributions can be identified:

- In contrast with traditional DRL approaches that directly exploit the entire action domain, our proposed method searches for well-developed low-level trading strategies, and consequently it can significantly narrow the search space and avoid the low-value explorations. To the best of the our knowledge, this is the first time that this approach has been used to address the portfolio optimization problem.

- Apart from the traditional publicly available market information, the state in our proposed approach is augmented through market indicators based on expert domain knowledge. These market indicators can

make use of more diversified information from alternative viewpoints of the market environment, as well as additional high-level and robust information to improve asset allocation decisions that might not be adequately stated when using only raw observations.

- The proposed DRL hyper-heuristic framework is evaluated on five real-world problem instances and numerous experimental results demonstrate our proposed method can achieve notable performance gains compared with state-of-art trading strategies as well as the DRL baseline method.

Besides, DRL is a cutting-edge technique that can be applied in financial trading and portfolio construction. Furthermore, portfolio construction is a high stake decision for financial institutions and corporations, indicating that a reliable technical approach for constructing portfolios serves as one of the key technological factors in changing financial markets in the future. As a result, how to adopt such a technique in a credible way could be extraordinarily attractive for the audience concerning future technological evolution in financial markets. Furthermore, technological innovations drive significant disruptions across industries, creating both challenges and opportunities for investors. Our multi-period portfolio optimization study can shed light on the impact of technological changes on investment decision-making. For instance, the rise of artificial intelligence and automation has transformed traditional industries, resulting in new investment opportunities in sectors such as machine learning and cybersecurity. By leveraging insights from our multi-period optimization models, investors can identify optimal asset allocations and risk management strategies to capitalize on these emerging trends.

Furthermore, our DRL based approach concentrates on the optimal allocation of assets over multiple periods, allowing investors to consider the dynamic nature of financial markets, the impact of changing economic conditions, and the need for rebalancing portfolios over time. As a result, our multi-period portfolio optimization model has significant policy implications for investors, financial institutions, and regulators. For example, our DRL-hyper heuristic framework can enrich regulatory policies by shedding light on the impact of regulations on portfolio decisions and market dynamics. For instance, based on our DRL-hyper heuristic framework, policy-makers can undertake the sensitivity analysis by examining the effect of capital requirements, margin regulations, or transaction taxes on portfolio optimization strategies. By understanding the implications of such regulations, policymakers can design more effective and efficient regulatory frameworks.

Our application of DRL hyper-heuristic to multi-period portfolio optimization, offers several advantages over existing methods. Firstly, this method can narrow the search space by searching for low-level trading strategies, which can help improve asset allocation decisions in real-world financial markets. Additionally, the proposed framework includes market indicators based on expert domain knowledge, further purifying the original data, to enhance the investment decision-making process. By adopting a DRL hyper-heuristic approach, the method overcomes one of the limitations of trend-prediction-based algorithms. These algorithms heavily rely on prediction accuracy, which cannot ensure optimal portfolio decisions. The DRL hyper-heuristic approach, on the other hand, uses deep reinforcement learning to optimize portfolio constructions based on observed market dynamics.

Nevertheless, this method still suffers from limitations, one potential limitation could be the computational complexity of implementing a DRL hyper-heuristic approach, as training deep reinforcement learning models can require significant computational resources. Additionally, the effectiveness of the chosen method may depend on the availability and quality of market indicators based on expert domain knowledge and original data.

Additionally, regarding the utility and application of this paper, our DRL hyper-heuristic framework can enhance the Liability-Driven Investment (LDI). LDI aims to align an investment portfolio with the liabilities of an institution, such as future pension payments. Optimized portfolio constructions across multiple periods can help institutional investors match their assets and liabilities over time, considering changing interest rates, inflation, and other factors that liabilities are sensitive to. As a result, financial institutions can build a portfolio that will provide a reliable income stream for their pension payments, while also protecting their investments from market volatility.

The rest of the paper is organized as follows: related work is provided in Section 2. Section 3 gives the problem description. The proposed solution framework is presented in Section 4. The experimental results are reported in Section 5. The research implications are summarized in Section 6. Section 7 concludes the paper.

2. Related work

The basic MPT optimization model is essentially a parametric quadratic programming problem. As a result, there exists exact algorithms that can obtain optimal solutions for most particular data sets efficiently. However, the inclusion of additional real-world trading constraints considerably enhances the problem's complexity. The majority of the time, adding new restrictions will result in a nonconvex search space, making it impossible to apply accurate methodologies. Hence, many practitioners and researchers try to adopt different heuristic or metaheuristic optimization techniques to solve the problem. Schaerf (2002) explored the use of local search techniques, mainly Tabu Search (TS), to the problem. Different neighborhood relations were investigated. Computation results were given for the 5 general market instances involving up to 225 assets. Crama and Schyns (2003) introduced a Simulated Annealing (SA) method to the practical portfolio optimization problem. The real-world trading constraints like cardinality, turnover constraints and trading constraints (minimum trading size) were also considered. The constraints were handled by a penalty approach, which adds a penalty term to the objective function for each violated constraint. Chang et al. (2000) highlighted the different shapes of the efficient frontier in the presence of cardinality constraint. They also showed that certain portions of the efficient frontier are disconnected. Three heuristic algorithms, which are Genetic Algorithm (GA), TS and SA, were then presented to solve the cardinality constrained model. Computational results were reported for 5 test instances (which are later made publicly available via the OR-library (Beasley, 1990) and used as the benchmark data sets) involving up to 225 assets. Fernández and Gómez (2007) applied a heuristic method based on artificial neural networks. Computational results were reported for the benchmark data set from the OR-library. They compared the results obtained by the artificial neural networks with the results of three heuristic algorithms reported in Chang et al. (2000) and concluded that no one heuristic outperformed the others in all kinds of investment policies. Chang et al. (2009) investigated GA to cardinality constrained portfolio optimization model with different risk measures. Computational results were reported for 3 test problems involving up to 99 assets. The authors also verified that investors should only consider one third of total assets to be selected into the portfolio since they are dominated by those contained more assets. Woodside-Oriakhi et al. (2011) presented 3 metaheuristic algorithms based on GA, TS and SA. The proposed metaheuristics make use of the subset optimization step in the sense that a (small) mixed-integer quadratic optimization problem can be solved to optimality. Better quality solutions were presented for the benchmark data set from the OR-library and it indicated that the subset optimization step is a useful strategy for the cardinality constrained portfolio optimization problem. Cura (2009) applied a Particle Swarm Optimization (PSO) approach where each particle represents a portfolio. Computational results were reported for the benchmark data set from the OR-library. They compared the results obtained by PSO with those obtained by GA, TS and SA and showed that none of the four algorithms could outperform the others in all kinds of investment policies. They also showed that PSO could obtain better solutions for the portfolio with a low risk level.

The applications of DRL in single asset trading have been witnessed in the recent literature, including critic-only method (Jeong and Kim, 2019), actor-only method (Deng et al., 2017; Wu et al., 2020) and actor-critic method (Li et al., 2018). As for portfolio construction that involves multiple assets, Almahdi and Yang (2017) utilized a recurrent RL approach to determine asset allocation weights and buy and sell signals using a coherent risk-adjusted performance objective function. Jiang et al. (2017) proposed an extensible RL framework based on Deep Deterministic Policy Gradient (DDPG) (Lillicrap et al., 2015) for optimizing cryptocurrency portfolios dynamically. Similarly, Cui et al. (2023) proposed a Conditional Value at Risk (CVaR) based DRL approach and used Proximal Policy Optimization (PPO) (Schulman et al., 2017) to construct portfolios in cryptocurrency market. Buehler et al. (2019) introduced a DRL framework for hedging a portfolio of derivatives in the face of different constraints. Ye et al. (2020) proposed a State Augmented RL (SARL) paradigm that incorporates asset information with price movement predictions as extra states in order to increase resilience against environmental unpredictability. Shi et al. (2022) implemented an RL framework based on graph convolutional networks that uses Relational Graph Convolutional Networks (R-GCN) to extract asset relational information and then integrates relational features with multi-scale temporal features to decide how to allocate assets. However, those approaches directly optimize asset weights without considering expert domain knowledge, thus exhibit poor scalability and quickly become intractable in real-world large problem instances.

3. Problem description

In this work, we consider a multi-period portfolio optimization problem. The asset allocation decisions are time-driven, meaning the weights of different assets within the portfolio need to be adjusted periodically. Specifically, the portfolio consists of 1 riskless asset (i.e., cash, as typically capital is not always fully invested, indexed by 0) and n financial assets (indexed by $1 \dots n$). The investment decisions on a trading period t are represented by a portfolio weight vector $\mathbf{w}^t = [w_0^t, w_1^t, w_2^t, \dots, w_n^t]^\intercal$, where w_i^t represents the proportion of total capital invested in the ith asset and $\sum_{i=1}^{n} w_i^t = 1$, $w_i^t \in [0,1]$, $t \in [1,T]$, T is the total trading duration. Let $\mathbf{o}^t = [o_0^t, o_1^t, o_2^t, \dots, o_n^t]^\intercal$, o_i^t denote the open price of the ith asset at time t, and $o_0^t = 1$ is the cash price. The initial portfolio weight vector $\mathbf{w}^1 = [1,0,0,\dots,0]^\intercal$, indicating that it purely consists of the riskless asset (cash), and the initial portfolio value, denoted by V^1, is the initial capital. Without loss of generality, the initial capital V^1 is set to be 1. The share holdings of asset i at time t, denoted by h_i^t, can be calculated as $\lfloor \frac{V^t * w_i^t}{o_i^t} \rfloor$, where V^t is the portfolio value at time t and $V^t = \sum_{i=1}^{n} h_i^{t-1} * o_i^t$, $t > 1$. When rebalancing the portfolio, transaction cost is applied and it is calculated as $\delta^t = \sum_{i=1}^{n} \Delta w_i^t * V^t * \beta$, where $\Delta w_i^t = |w_i^t - w_i^{t-1}|$ is the adjusted proportion of the ith asset, β is a constant commission rate, $t > 1$. The portfolio return at time t ($t > 1$), denoted by r^t, can be calculated as $((V^t - \delta^t) - (V^{t-1} - \delta^{t-1}))/(V^{t-1} - \delta^{t-1})$. The portfolio risk at time t ($t > 1$), is calculated as $\sum_{i=1}^{n} \sum_{j=1}^{n} w_i^t w_j^t \sigma_{ij}^t$ where σ_{ij}^t is the covariance between assets i and asset j ($i, j \in [1,n]$) at time t. Specifically, the covariance matrix $\sigma_{n \times n}$ is symmetric and each diagonal element σ_{ii} represents the variance of the asset i, while the covariance σ_{ij} represents the correlated risks between asset i and asset j. The total return r^T at the end of the trading duration T can be calculated as the accumulated product of the portfolio return $\prod_{t=1}^{T} r^t$. The objective is to maximize the accumulative portfolio return for a given time frame while minimizing the risk at the same time.

For this study, backtesting trading settings are applied and for simplicity, the imposed presumptions (Jiang et al., 2017) are as follows:

- Zero slippage: all market assets have sufficient liquidity, allowing any order to be executed right away.
- Insignificant impact: the agent's investing decisions will not have an impact on the market.
- Asset liquidity: The volume of each asset is large enough, and as a result, the agent can buy or sell any asset at any trading period.

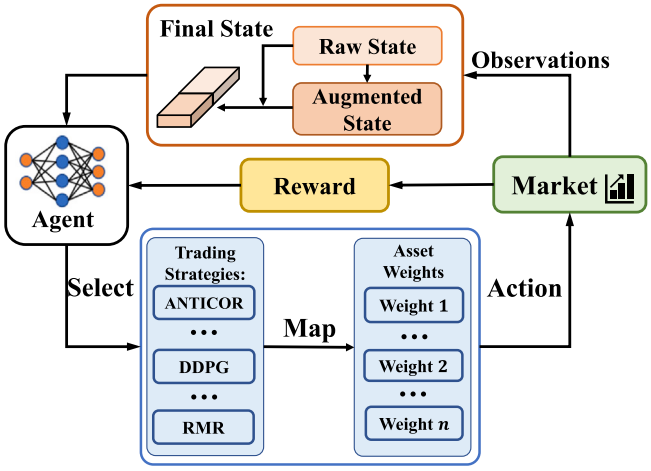

Fig. 1. The proposed DRL-based hyper-heuristic framework.

4. Our proposed solution framework

The examined problem is formulated as a Markov Decision Process (MDP): $(\mathcal{S}, \mathcal{A}, \mathcal{P}_a, R_a, \gamma)$ where $\mathcal{S}$ is the state space, $\mathcal{A}$ is the action space, $\mathcal{P}_a(s,s') = Pr(s_{\tau+1} = s' | s_\tau = s, a_\tau = a)$ is the probability that action a in state s at time instance τ will lead to state s' at next time instance $\tau+1$, $R_a(s,s')$ is the instant reward received after transitioning from state s to state s', due to action a and $\gamma \in [0,1)$ is the discount factor. The agent's goal is to explore a policy π that can maximize the cumulative discounted reward as calculated by $\sum_{\tau=0}^{T} \gamma^\tau R_{a,\tau}$.

The proposed solution framework is depicted in Fig. 1. Note that our framework is compatible with standard DRL algorithms and the choices are not limited. In this work, we utilize a standard on-policy DRL algorithm, PPO, to train a heuristic selection module from a set of low-level trading strategies (which can be considered low-level heuristics in hyper-heuristics framework) in different capital markets. PPO computes an update at each step that optimizes the cost function while maintaining the minor variations between the new and old policies, necessitating little parameter tuning work. The selection of low-level trading strategies is based on two state vectors as well as the experiences of the DRL agent. Specifically, our proposed approach focuses on both expert domain knowledge (via market indicator states) and market raw information (via market states). Instead of searching for the asset weights directly, our proposed method searches for well-developed trading strategies, which can be further mapped to asset weights accordingly. In such circumstances, it tends to be easier to train the DRL agent compared with previous approaches (Jiang et al., 2017; Ye et al., 2020; Cui et al., 2023) due to the use of additional information and the reduced search space. Moreover, the low-level trading strategies encapsulate subjective perceptions and expectations into actionable simple rules, thus they are more robust in the presence of uncertainties (Gilbert-Saad et al., 2023).

4.1. State

The state features of our proposed framework are illustrated in Fig. 2. For a given trading period t, there are four points that can be used to capture the overall price movement, namely, opening($\mathbf{o}^t$), high($\mathbf{h}^t$), low($\mathbf{l}^t$) and closing($\mathbf{c}^t$) (Rogers and Satchell, 1991), where $\mathbf{o}^t = [o_0^t, \dots, o_n^t]^\intercal$, $\mathbf{h}^t = [h_0^t \dots, h_n^t]^\intercal$, $\mathbf{l}^t = [l_0^t, \dots, l_n^t]^\intercal$, $\mathbf{c}^t = [c_0^t, \dots, c_n^t]^\intercal$, n is the total number of assets. These four price vectors, along with the trading volume $\mathbf{v}^t = [v_0^t, \dots, v_n^t]^\intercal$, compose the exogenous state features which are decoupled from the agent's actions. The current portfolio value V^t and the current portfolio return r^t compose the endogenous state features that are affected by the agent's actions. Those exogenous and

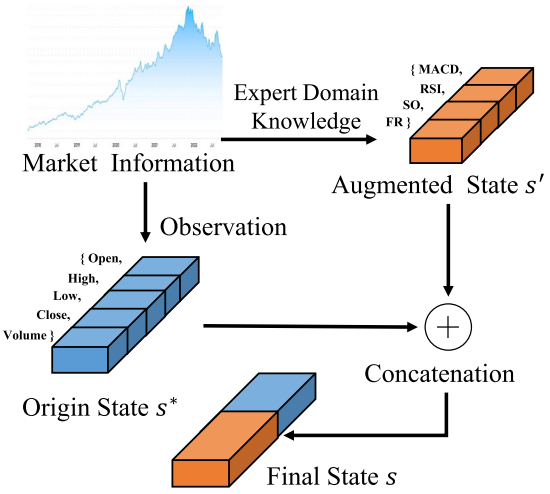

Fig. 2. State features of proposed DRL hyper-heuristic framework.

endogenous state features constitute the origin state s_t^* of the examined problem.

However, the origin state may contain a high degree of noise and uncertainty, plus non-stationary trait and therefore cause a distribution shift between training and testing data. To tackle this, market technical indicators based on expert domain knowledge are introduced to summarize markets' behavior from different perspectives and extract useful patterns. These indicators are heuristics or mathematical estimations based on the general raw market data. Some popular technical indicators include Moving Average Convergence Divergence (MACD) (Appel, 2005), Relative Strength Index (RSI) (Wilder, 1978), Stochastic Oscillator (SO) (Cao et al., 2020) and Fibonacci Retracement (FR) (Tsinaslanidis et al., 2022). In specific terms, MACD is a trend-following momentum indicator based on two moving average lines of a particular asset price. The difference between long-run and short-run trends can be captured by MACD, which can be used as a trend follower. Similarly, RSI is also a momentum indicator that quantifies the magnitude of price changes, suggesting the overbought or oversold conditions for a particular asset. SO also measures the overbought or oversold conditions by fully considering the random amplitude of price fluctuations and the measurement of short-run and medium-run fluctuations in the design, making its short-term market measurement function more accurate and effective. FR generates the impulse line from each golden section point (0.618) against the current stock price. Those lines could be extremely helpful to identify areas where buyers may be accumulating heavy buying pressure after the price drop.

To further leverage additional diverse information, we also include the aforementioned four market indicator vectors, $\boldsymbol{\alpha}^t = [\alpha_0^t, \ldots, \alpha_n^t]^\intercal$, $\boldsymbol{\eta}^t = [\eta_0^t, \ldots, \eta_n^t]^\intercal$, $\boldsymbol{\iota}^t = [\iota_0^t, \ldots, \iota_n^t]^\intercal$, $\boldsymbol{\xi}^t = [\xi_0^t, \ldots, \xi_n^t]^\intercal$ as the augmented state s_t' where $\boldsymbol{\alpha}$ is MACD, $\boldsymbol{\eta}$ is RSI, $\boldsymbol{\iota}$ is SO, and $\boldsymbol{\xi}$ is FR, respectively.

Thus, the final state $\boldsymbol{s}_t = [\mathbf{o}^t, \mathbf{h}^t, \mathbf{l}^t, \mathbf{c}^t, \mathbf{v}^t, \boldsymbol{\alpha}^t, \boldsymbol{\eta}^t, \boldsymbol{\iota}^t, \boldsymbol{\xi}^t, r^t, V^t]$ at a single time step t is the concatenation of origin state s_t^* and augmented state s_t' with the dimension of $9n + 2$. In this work, some particular actions (trading strategies) may require a certain time window for executing, thus we consider a time instance τ consisting of m time steps t and therefore the state s_τ at time instance τ is an m by $(9n+2)$ matrix.

4.2. Action

Normally, in hyper-heuristic framework, the actions are various heuristic rules. In our proposed solution framework, there are two levels of actions. Given the state s_τ, the agent may firstly perform an agent action a_τ to select a sophisticated trading strategy. The action space is defined as a 12-dimensional vector $\mathcal{A} = [a_1, \ldots, a_{12}]$ where each element represents one specific trading strategy. Specifically, we consider the eleven most prevailing trading strategies (Li et al., 2016), namely, Anti-correlation strategy (Anticor), Buy and Hold strategy (BAH), Best Constant Rebalanced Portfolio strategy (BCRP), Nonparametric Nearest Neighbor log-optimal trading strategy (BNN), Constant Rebalanced Portfolio strategy (CRP), Confidence Weighted Mean Reversion (CWMR), Exponential Gradient (EG), On-Line Moving Average Reversion strategy (OLMAR), Online Newton Step (ONS), Passive Aggressive Mean Reversion strategy (PAMR), Robust Median Reversion strategy (RMR) as well as one baseline DRL trading strategy using DDPG (Jiang et al., 2017). A detailed description can be found in Table 1. Once the agent action a_τ is selected, the assets' weights at time instance τ can be obtained instantly by simply executing the corresponding trading strategy. Our proposed framework can be easily extended by incorporating more actions (trading strategies).

Table 1
Brief description of the trading strategies (agent actions) used in proposed DRL hyper-heuristic framework.

Agent action	Brief description
Anticor	Portfolio construction based on stock correlations and anti-correlations in consecutive windows.
BAH	Buy and hold the asset selected, retaining the investment ignoring short-term ups and downs in market price.
BCRP	Redistribute the investment wealth each trading day based on hindsight.
BNN	A nearest neighbor-based strategy exploited by histograms from the nonparametric statistics method.
CRP	On a daily basis, maintain the same wealth distribution among a particular group of assets.
CWMR	Create a Gaussian distribution to represent the portfolio vector, and then update it in accordance with the mean reversion principle.
DDPG	Baseline DRL strategy for constructing portfolios.
EG	Track the best stock and adopt regularization term to reduce departure from the prior portfolio.
OLMAR	Forecast next price relatives based on the moving average method and then construct portfolios via online learning techniques.
ONS	Track the best CRP to date and adopt a L2-norm regularization to limit the variability of portfolio.
PAMR	Adopt the mean reversion model of financial time series based on online passive aggressive learning.
RMR	Construct portfolios based on the median reversion property of financial time series using a robust L1-estimator and passive aggressive online learning.

4.3. State transition

The state transition from s_τ to $s_{\tau+1}$ is governed by the function $s_{\tau+1} = F(s_\tau, a_\tau, \varphi_\tau)$. The transition may not only depend on the action a_τ but can also be affected by the uncertainties φ_τ that exist in some state features. In this work, the transitions for O^τ (m by n opening price matrix), H^τ (m by n high price matrix), L^τ (m by n low price matrix), C^τ (m by n closing price matrix) and v^τ (m by n volume matrix) are subject to market uncertainties. On the other hand, the transitions for $\mathbf{r}^\tau$ (m-dimensional vector) and $\mathbf{V}^\tau$ (m-dimensional vector) are directly affected by the agent's actions. For a given time instance τ, the portfolio is rebalanced after action a_τ is executed.

4.4. Reward

The immediate reward R_τ resulting from the agent's action a_τ is set to be equal to portfolio return minus portfolio risk, as calculated by $r^\tau - \sum_{i=1}^{n}\sum_{j=1}^{n} w_i^\tau w_j^\tau \sigma_{ij}^\tau$.

5. Experimental results

5.1. Data sets and experimental settings

For this work, we choose the following five different real-world capital markets suggested by OR-library (Beasley, 1990) to evaluate the performance of our proposed solution framework for multi-period portfolio optimization problem:

- DAX in Germany, total number of assets $n = 33$.
- Hang Seng in Hong Kong, total number of assets $n = 55$.
- FTSE in UK, total number of assets $n = 88$.
- S&P in US, total number of assets $n = 95$.
- Nikkei in Japan, total number of assets $n = 152$.

The data sampling period is from 2012-09-26 to 2022-07-22. For each market instance, we split the data set into 80% training and 20% testing. The single trading period t is set to be 1 day and the action selection time window m is set to be 10. For simplicity, the constant commission rate $\beta = 0.05\%$ is applied for all market instances.

5.2. Evaluation

We contrast the effectiveness of the portfolio constructed by the proposed DRL hyper-heuristic framework with each portfolio construction based on the existing state-of-art trading strategies listed in Table 1, as well as the standard PPO method (PPO searching for low-level action directly) and general market trend (market index).

We adopt three widely-used metrics for evaluation. The final portfolio value V^T is used to show how asset values have changed cumulatively over the trading duration T. The Annualized Return (AR), which is based on the portfolio value, can also be used to show the geometric average of a portfolio's annual earnings. Specifically, AR can be calculated as $(1 + r^T)^{365/T} - 1$. Since both portfolio value and AR cannot provide investors any indication of the portfolio risk, Sharpe Ratio (SR) (Sharpe, 1994), is also used to simultaneously evaluate both profit and risk of constructed portfolios. Conceptually, Sharpe Ratio is utilized to calculate the predicted return per unit of risk and it can be computed as $(r_p - r_f)/\rho_p$ where r_p is the portfolio return, r_f is the risk-free rate and ρ_p is the standard deviation of the portfolio's excess return. In this work, $r_p = r^T$ for a trading duration T and $r_f = 3\%$ is a typical bank interest value.

Our model's performance can yield theoretical, empirical and marginal economic benefits. Our research implications can yield theoretical, empirical and marginal economic benefits. First, our study of multi-period portfolio optimization using DRL can contribute to the theoretical understanding of reinforcement learning algorithms and their application in financial decision-making. The development of DRL hyper-heuristic based policies that can capture the time-varying nature of risk and return. Our DRL hyper-heuristic framework thereby can adapt to changing market conditions, detect patterns, and exploit market inefficiencies. Our model can enhance the understanding of the dynamics of financial markets and provide insights into optimal investment strategies over multiple periods. This paper thereby enhances the theoretical knowledge of how the stock market might evolve and then how to effectively construct optimal portfolio in such a real-world investment scenario. Empirically, our DRL hyper-heuristic framework on multi-period portfolio optimization in real stock markets provides valuable insights into the performance and effectiveness of this approach. By applying our framework to different historical market data, investors can evaluate their ability to generate superior risk-adjusted returns compared to traditional portfolio optimization methods as well as the stock market indices. Finally, the marginal economic benefit that our model can bring rely on the improved risk-adjusted returns. Our DRL based approach can help investors to create superior risk-adjusted returns by capturing complex patterns and adapting to changing market conditions. The application of our framework can lead to better portfolio performance and enhance long-term investment outcomes for investors from economic benefit perspective.

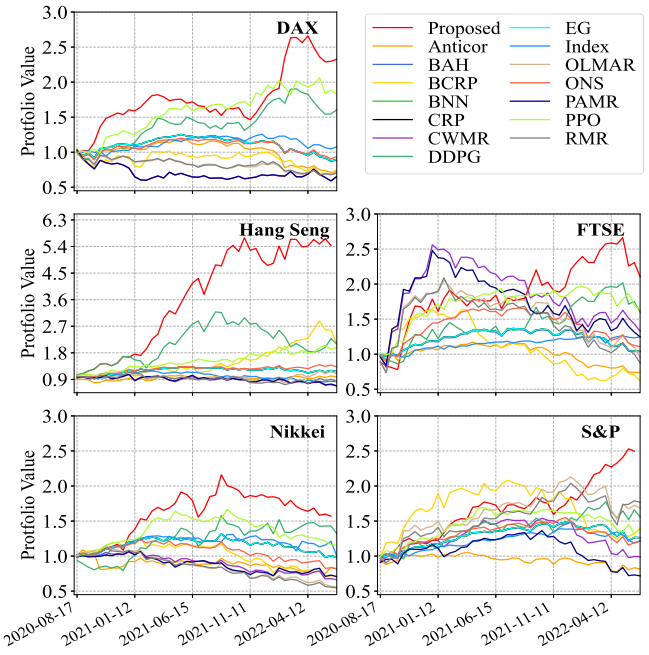

Fig. 3. Testing results for portfolio value of different methods on five real-world market instances.

5.3. Results and discussions

5.3.1. Comparison of performance on five real-world market instances over back-testing period

The proposed DRL hyper-heuristic method is trained for 3000 episodes with 10 different random seeds. During the training, each episode represents one trading period between 2012-09-26 and 2020-08-16. After the training, the trained actor is deployed to select different trading strategies between 2020-08-17 and 2022-07-22 for performance evaluation.

Our proposed approach is benchmarked against 13 state-of-art trading strategies as well as the general market trend (market index). The portfolio values for the back-testing period on 5 real-world market instances are shown in Fig. 3 while the annualized return and Sharpe ratio are reported in Table 2.

Fig. 3 demonstrates that our proposed method has the dominating performance on 3 (DAX, Hang Seng & Nikkei) out of 5 instances and outperforms the other trading strategies on all 5 market instances over the entire testing period. In terms of AR, our proposed method overwhelmingly outperforms all other algorithms, improving the second-best method by 65%, 296%, 75%, 78% and 94% on DAX, Hang Seng, FTSE, Nikkei and S&P, respectively. Normally, in multi-period portfolio optimization, SR often decreases over the back-testing period due to the environmental distribution shift. In this work, we penalize the portfolio risk in the reward setting and our proposed method can obtain the highest SR compared with other trading strategies. In practice, a SR higher than 2.0 is rated as very good and our proposed method can attain 2.0+ SR on all 5 market instances, indicating the risk is paying off in the form of above-average returns.

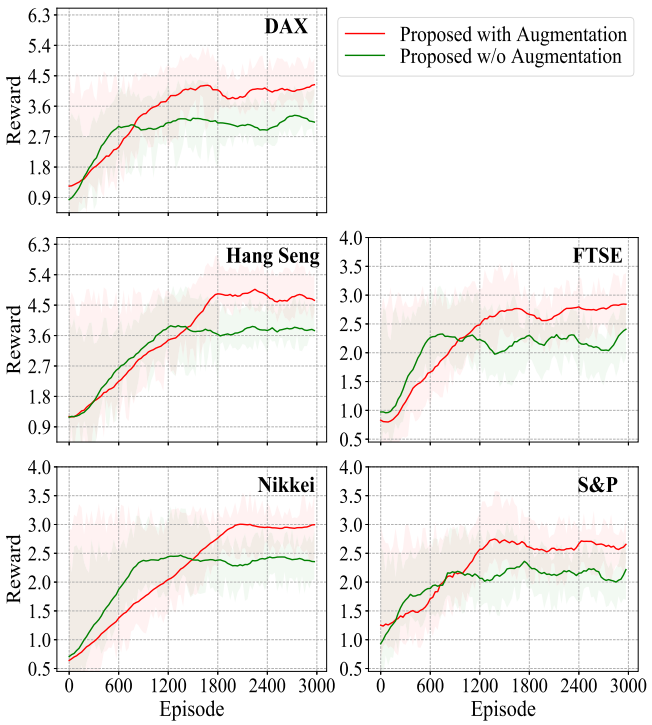

Fig. 4. Episodic moving average reward of proposed method with and without augmentation on five real-world market instances.

Table 2
Annualized return and Sharpe ratio of different methods on five real-world market instances.

Method	Instance		DAX	Hang Seng	FTSE	Nikkei	S&P
Proposed		AR(%)	**73.33**	**246.11**	**61.11**	**31.61**	**82.78**
		SR	**2.49**	**3.45**	**2.11**	**2.02**	**2.33**
Anticor		AR(%)	−15.11	0.81	−14.56	−10.12	−10.19
		SR	−1.3	−0.65	−1.34	−1.83	−1.95
BAH		AR(%)	−6.11	8.56	1.9	1.29	14.48
		SR	−0.84	0.59	−0.09	−0.48	0.82
BCRP		AR(%)	−13.44	62.22	28	−15.38	21.83
		SR	−1.76	1.15	−0.72	−1.16	0.62
BNN		AR(%)	−5.67	9.01	2.28	−1.19	14.7
		SR	−0.81	0.65	−0.06	−0.47	0.83
CRP		AR(%)	−6.22	9.01	2.28	−1.19	14.7
		SR	−0.81	0.65	−0.06	−0.47	0.83
CWMR		AR(%)	−19.33	−18.3	18.22	−18.77	−0.45
		SR	−2.38	−2.47	0.37	−1.8	−0.18
DDPG		AR(%)	33.33	60.56	33.33	17.78	31.11
		SR	1.13	0.97	1.04	0.62	1.34
EG		AR(%)	−5.67	9.04	2.28	−0.12	14.69
		SR	−0.82	0.65	−0.06	−0.47	0.83
Index		AR(%)	4.29	−9.86	13.76	11.65	12.32
		SR	0.152	−1.21	1.28	0.98	0.77
OLMAR		AR(%)	−15.94	4.21	−2.61	−24.44	38.09
		SR	−2.4	0.15	−0.17	−1.62	1.17
ONS		AR(%)	−3.5	18.32	6	−10.11	11.58
		SR	−0.74	1.42	0.14	−1.17	0.61
PAMR		AR(%)	−19.22	−17.43	13.72	−16.34	−15.64
		SR	−2.37	−2.17	0.28	−1.8	0.28
PPO		AR(%)	44.44	54.3	35.00	8.33	22.22
		SR	1.38	1.79	1.24	0.30	0.99
RMR		AR(%)	−16.61	−9.41	−6.89	−25.25	42.7
		SR	−2.44	−1.3	−0.29	−1.65	1.42

To validate the effectiveness of incorporating expert domain knowledge, we compare the training process between the proposed method with and without state augmentation. The results are illustrated in Fig. 4. It is evident that the learning curves of proposed method with raw states grow relatively fast, but converge at lower rewards. The DRL hyper-heuristic approach with augmented state shows superior performances on all market instances, demonstrating the importance of using expert's guidance to gain deeper insights into price movements that may not be sufficiently expressed when merely using assets' intrinsic value based on messy raw market information.

5.3.2. Portfolio composition

In practice, diversification is a common principle that entails buying different assets to mitigate overall portfolio risk. Empirically, diversification can increase portfolio returns and reduce portfolio risk, and also provide higher returns as compared to the traditional portfolios for the same level of risk (Ma et al., 2020). In practice, rebalancing is a mechanism for replenishing diversification to establish better risk control. We compare the portfolio compositions of our proposed method with baseline DRL strategies (DDPG and PPO) that directly operate in the low-level action space for the last 50 days in the testing period, as shown in Fig. 5. It can be found that the portfolios obtained by our method are well diversified and thus more robust since they are composed of a suitable number of assets. In contrast, the portfolios obtained by DDPG tend to become more concentrated in the assets that have previously performed well, which are fragile in the face of unintentional exposure risk, while the portfolios obtained by PPO are too cluttered, which may lead to high transaction costs when rebalancing.

5.3.3. Action distribution

To further understand the trained policy, the action distributions are summarized for all 5 market instances in the testing period, as shown in Fig. 6. It is worth mentioned that although some strategies fail to earn the positive profit over the entire testing period, their short-term performance might still be favored. For example, in DAX, ONS constitutes 37.0% of the total actions while its AR is only −3.5%. BCRP and EG also obtain negative ARs, but still constitutes 19.6% and 13.0% of the total actions, respectively. The findings show that there are patterns between particular market circumstances and the related strategies. In practice, there is an overwhelming number of options for choosing different trading strategies. Our proposed DRL hyper-heuristic approach can provide an effective framework for how to use them in combination.

5.3.4. Robustness test

For the robustness purpose, we have compared our results with different stock market indices. As stated in the efficient market theory, stock market participants can instantaneously correct the market mispricing, suggesting that beating the market would be unlikely. Therefore, we compared our trading results with corresponding stock market indices. However, we demonstrate that the proposed DRL hyper-heuristic framework can be used to beat the market performance by comparing with those stock market indices. In particular, our DRL hyper-heuristic framework had the annualized return of 73.33%, 246.11%, 61.11%, 31.61% and 82.78%, for five stock markets respectively, and on the other side, the performance of corresponding stock market indices were 4.29%, −9.86%, 13.76%, 11.65%, 12.32%. Therefore, our results exhibit novelties such that our DRL hyper-heuristic framework can beat the market indices in all five countries, complementing the efficient market theory that market participants cannot discover the market mispricing in a prompt manner and the algorithm can thereby grasp those investment opportunities.

6. Research implications

Our multi-period portfolio optimization based on DRL hyper-heuristic framework can deliver fruitful research implications. Firstly, the multi-period portfolio optimization model can align with the Intertemporal Capital Asset Pricing Model (ICAPM) proposed by Merton (1973). On the basis, our the multi-period portfolio optimization model

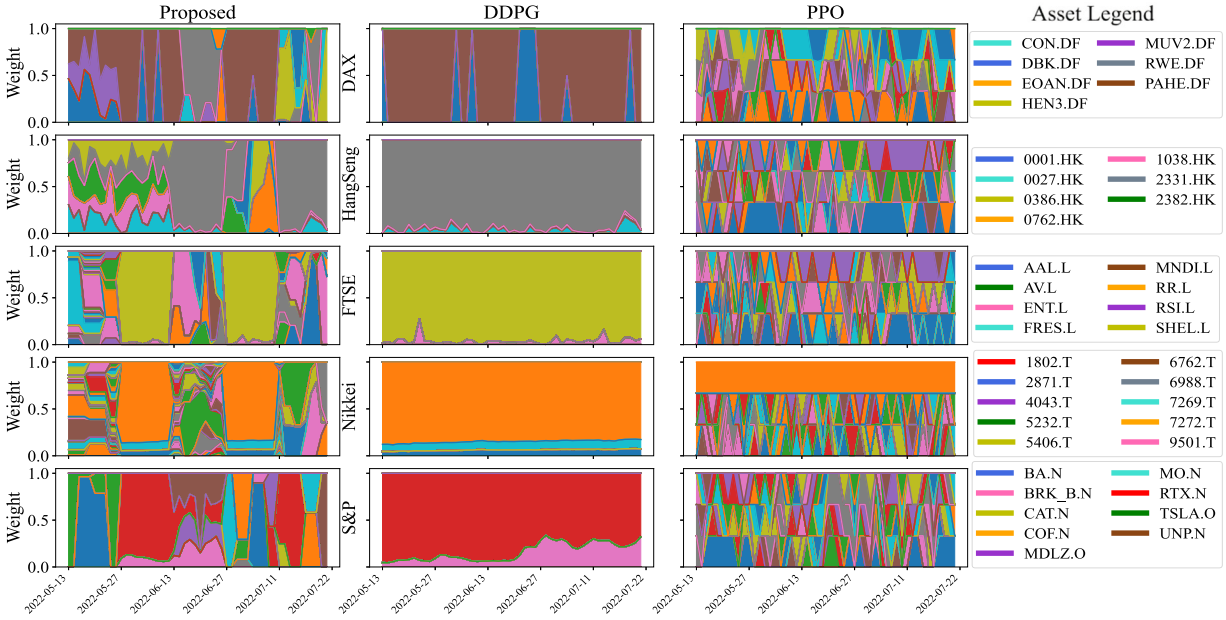

Fig. 5. Comparison of portfolio compositions between proposed method and DRL baselines (DDPG, PPO) on five real-world market instances.

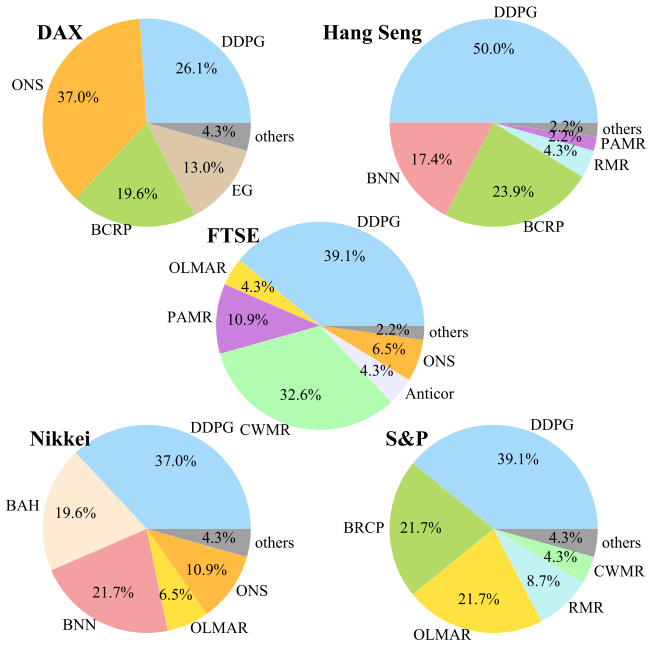

Fig. 6. The distribution of the actions of the proposed method in testing period on five real-world market instances.

enables long-term investors to adjust their portfolio holdings in a timely manner to frustrate the deterioration of stock performance as the investment environment evolves (Campbell et al., 2018). Multi-period portfolio construction significantly differs from the single-period portfolio construction as multi-period portfolio construction may face more complex market conditions than single-period portfolio construction. As a result, multi-period portfolio construction typically adopts a dynamic investment strategy that adjusts the portfolio allocation over time in response to changing market conditions and evolving investor circumstances. This strategy allows for rebalancing the portfolio and taking advantage of opportunities that arise. Single-period portfolio construction, in contrast, often adopts a static investment strategy, where the initial portfolio allocation is maintained throughout the specified period without significant adjustments.

Furthermore, our model is also helpful in constructing the time-consistent portfolio since our model can dynamically rebalance the portfolio holding to retain the optimum in a multi-period time scale. The key prerequisite for time-consistent portfolios is that they should be dynamically optimal (Pun, 2018; Peng and Kloeden, 2021). Compared with the single-period portfolio optimization model, our multi-period model is superior in addressing plenty of portfolio management issues, including changes in market dynamics and goal-based risk measures (Li et al., 2022). Finally, our model incorporates additional diverse information, which allows investors to implement high-dimensional data analysis. High dimensional covariance of market indicators could be a pivotal constituent for optimal asset allocation decisions (Hautsch et al., 2015). Nevertheless, the performance of traditional approaches with high-dimensional data could be unreliable since they usually underestimate the risk, resulting in suboptimal results (Bodnar et al., 2018). Therefore, our model can aid the model performance with efficient analysis of high-dimensional data analysis based on the proposed framework, delivering an eminent portfolio construction strategy in a multi-period time horizon, and can be further embedded in robo advisors (Tao et al., 2021) for efficient investment management.

Currently, in the context of the Russia–Ukraine conflicts and COVID-19, geopolitical risks and regulatory uncertainties can significantly impact the performance of portfolios. Our DRL hyper-heuristic framework can provide investors to construction an optimized portfolio. Through the optimized portfolio construction, investors can proactively manage these risks by hedging currency exposures, employing derivatives to protect against adverse market movements, and adjusting portfolio allocations based on the evolving geopolitical situation. During the COVID-19 pandemic, optimized portfolio constructions could have incorporated risk management techniques to mitigate the impact of market downturns and volatility. This could include the use of defensive assets, such as government bonds or gold, to provide a hedge against equity market declines. Additionally, strategies like portfolio insurance, where the allocation to defensive assets is dynamically adjusted based on predefined risk triggers, could have helped limit downside risk during the sharp market sell-offs. As the pandemic and Russia–Ukraine conflicts situation continue to evolve, investors may need to adjust their portfolios to reflect changing market conditions and investor preferences. By adopting our DRL hyper-heuristic framework

to continuously analyze new data and adjust the weights of different assets in the portfolio, investors can ensure that their portfolio remains well-diversified and optimized for the specific risk and return objectives of the investor.

7. Conclusion

This paper showcases the first application of DRL hyper-heuristic framework to multi-period portfolio optimization problem. By taking the advantage of well-developed low-level trading strategies, the DRL agent can effectively narrow the action space and improve the overall solution quality. Moreover, a state augmentation scheme based on expert domain knowledge is utilized to further improve the performance of the proposed method.

The results of the experiments show that our proposed framework has a number of benefits. Firstly, it shows better performance compared with state-of-art trading strategies as well as DRL baseline method that directly exploit the low-level action space. Secondly, it can provide more diversified portfolio constructions and thus enhance robustness against market uncertainties. Finally, the proposed algorithmic framework reveal that certain patterns exist between market conditions and trading strategies, allowing investors more easily understand and accept the suggestions provided by the proposed method.

Our study can exhibit salient policy implications. Our DRL hyper-heuristic framework can enrich regulatory policies by shedding light on the impact of regulations on portfolio decisions and market dynamics. In addition, regarding the utility and application of this paper, our DRL hyper-heuristic framework can strengthen the LDI, such as the pension plan investments. Our study of multi-period portfolio optimization using DRL hyper-heuristic in real stock markets can also create various benefits to different stakeholders. These stakeholders include individual investors, asset managers, financial institutions, regulators. For example, an asset management firm that implements our DRL hyper-heuristic portfolio optimization model may outperform competitors in terms of risk-adjusted returns and attract a larger client base seeking superior investment performance. Our study can also contribute to the theoretical understanding of RL algorithms and their application in financial decision-making. The development of DRL based portfolio optimization models that can capture the time-varying nature of risk and return. Our paper thereby strengthens the theoretical understanding of the financial market dynamics as well as optimal portfolio construction by applying DRL in a real-world investment scenario.

The relevance of our paper to the financial research field lies in its ability to address the dynamic nature of financial markets and the need for rebalancing portfolios over time. It has significant policy implications for investors, financial institutions, and regulators, as mentioned before. Additionally, our paper contributes to the theoretical understanding of reinforcement learning algorithms and their application in financial decision-making. This understanding further enhances Liability-Driven Investment and aligns portfolio constructions with future liabilities such as pension payments by using the DRL based method.

Our paper further pushes the boundaries of existing theoretical paradigms by introducing a novel DRL hyper-heuristic approach to multi-period portfolio optimization problems in real-world financial markets. The proposed approach goes beyond traditional DRL methods by searching for well-developed low-level trading strategies instead of directly exploiting the entire action domain. Additionally, the paper incorporates market indicators based on expert domain knowledge to augment the state and provide additional high-level and robust information for improving asset allocation decisions.

Furthermore, the impacts of our study are also remarkable. Our DRL hyper-heuristic framework can continuously learn from historical market data, monitor portfolio performance, and dynamically adjust allocations to manage risk. Investors can implement adaptive risk management strategies by applying our hyper-heuristic framework. This adaptive risk management approach allows investors to respond more effectively to changing market conditions, reduce downside risk, and potentially limit losses during periods of market stress or volatility, which can impact the stock market trading behavior.

CRediT authorship contribution statement

Tianxiang Cui: Conceptualization, Methodology, Investigation, Writing – original draft, Writing – review & editing, Validation, Formal analysis, Supervision. **Nanjiang Du:** Validation, Formal analysis, Data curation, Investigation, Methodology, Visualization. **Xiaoying Yang:** Validation, Formal analysis, Data curation, Investigation, Methodology, Visualization. **Shusheng Ding:** Validation, Investigation, Visualization, Writing – review & editing.

Data availability

Data will be made available on request.

Acknowledgment

This Project is Supported by Ningbo Natural Science Foundation, China (Project ID 2023J194), and by Ningbo Government, China (Project ID 2021B-008-C).

References

Abedin, M., Moon, M., Hassan, M., Hajek, P., 2021. Deep learning-based exchange rate prediction during the COVID–19. Ann. Oper. Res. This article was supported by the scientific research project of the Czech Sciences Foundation Grant No. 19-15498S.
Ahmed, L., Mumford, C., Kheiri, A., 2019. Solving urban transit route design problem using selection hyper-heuristics. European J. Oper. Res. 274 (2), 545–559.
Almahdi, S., Yang, S.Y., 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.
Ang, A., Bekaert, G., 2007. Stock return predictability: Is it there? Rev. Financ. Stud. 20 (3), 651–707.
Appel, G., 2005. Technical Analysis: Power Tools for Active Investors. FT Press.
Avramov, D., 2002. Stock return predictability and model uncertainty. J. Financ. Econ. 64 (3), 423–458.
Beasley, J., 1990. OR-library: distributing test problems by electronic mail. J. Oper. Res. Soc. 41 (11), 1069–1072.
Bellman, R., 1957. Dynamic Programming. Princeton University Press, Princeton, NJ.
Bertsimas, D., Shioda, R., 2009. Algorithm for cardinality-constrained quadratic optimization. Comput. Optim. Appl. 43 (1), 1–22.
Bodnar, T., Parolya, N., Schmid, W., 2018. Estimation of the global minimum variance portfolio in high dimensions. European J. Oper. Res. 266 (1), 371–390.
Bonami, P., Lejeune, M.A., 2009. An exact solution approach for portfolio optimization problems under stochastic and integer constraints. Oper. Res. 57 (3), 650–670.
Boubaker, S., Liu, Z., Zhai, L., 2021. Big data, news diversity and financial market crash. Technol. Forecast. Soc. Change 168, 120755.
Buehler, H., Gonon, L., Teichmann, J., Wood, B., 2019. Deep hedging. Quant. Finance 19 (8), 1271–1291.
Burke, E.K., Hyde, M.R., Kendall, G., Ochoa, G., Ozcan, E., Woodward, J.R., 2019. A classification of hyper-heuristic approaches: Revisited. In: Handbook of Metaheuristics. Springer, pp. 453–477.
Campbell, J.Y., Giglio, S., Polk, C., Turley, R., 2018. An intertemporal CAPM with stochastic volatility. J. Financ. Econ. 128 (2), 207–233.
Cao, A., Lindner, B., Thomas, P.J., 2020. A partial differential equation for the mean–return-time phase of planar stochastic oscillators. SIAM J. Appl. Math. 80 (1), 422–447.
Chang, T.J., Meade, N., Beasley, J.E., Sharaiha, Y.M., 2000. Heuristics for cardinality constrained portfolio optimisation. Comput. Oper. Res. 27 (13), 1271–1302.
Chang, T.-J., Yang, S.-C., Chang, K.-J., 2009. Portfolio optimization problems in different risk measures using genetic algorithm. Expert Syst. Appl. 36 (7), 10529–10537.
Chu, J., Zhang, Y., Chan, S., 2019. The adaptive market hypothesis in the high frequency cryptocurrency market. Int. Rev. Financ. Anal. 64, 221–231.
Crama, Y., Schyns, M., 2003. Simulated annealing for complex portfolio selection problems. European J. Oper. Res. 150 (3), 546–571.
Cui, T., Bai, R., Ding, S., Parkes, A.J., Qu, R., He, F., Li, J., 2020. A hybrid combinatorial approach to a two-stage stochastic portfolio optimization model with uncertain asset prices. Soft Comput. 24 (4), 2809–2831.

Cui, T., Cheng, S., Bai, R., 2014. A combinatorial algorithm for the cardinality constrained portfolio optimization problem. In: IEEE Congress on Evolutionary Computation. CEC, pp. 491–498.
Cui, T., Ding, S., Jin, H., Zhang, Y., 2023. Portfolio constructions in cryptocurrency market: A CVaR-based deep reinforcement learning approach. Econ. Model. 119, 106078.
Cura, T., 2009. Particle swarm optimization approach to portfolio optimization. Nonlinear Anal. RWA 10 (4), 2396–2406.
Deng, Y., Bao, F., Kong, Y., Ren, Z., Dai, Q., 2017. Deep direct reinforcement learning for financial signal representation and trading. IEEE Trans. Neural Netw. Learn. Syst. 28 (3), 653–664.
Ding, S., Cui, T., Bellotti, A.G., Abedin, M.Z., Lucey, B., 2023. The role of feature importance in predicting corporate financial distress in pre and post COVID periods: Evidence from China. Int. Rev. Financ. Anal. 90, 102851.
Eachempati, P., Srivastava, P.R., Kumar, A., Tan, K.H., Gupta, S., 2021. Validating the impact of accounting disclosures on stock market: A deep neural network approach. Technol. Forecast. Soc. Change 170, 120903.
Efat, M., Hajek, P., Abedin, M., Azad, R., Jaber, M., Aditya, S., Hassan, M., 2022. Deep-learning model using hybrid adaptive trend estimated series for modelling and forecasting sales. Ann. Oper. Res.
Fama, E.F., 1965. The behavior of stock-market prices. J. Bus. 38 (1), 34–105.
Fama, E.F., 1970. Efficient capital markets: A review of theory and empirical work. J. Finance 25 (2), 383–417.
Fernández, A., Gómez, S., 2007. Portfolio selection using neural networks. Comput. Oper. Res. 34 (4), 1177–1191.
Gilbert-Saad, A., Siedlok, F., McNaughton, R.B., 2023. Entrepreneurial heuristics: Making strategic decisions in highly uncertain environments. Technol. Forecast. Soc. Change 189, 122335.
Hautsch, N., Kyj, L.M., Malec, P., 2015. Do high-frequency data improve high-dimensional portfolio allocations? J. Appl. Econometrics 30 (2), 263–290.
Jeong, G., Kim, H.Y., 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.
Jiang, Z., Xu, D., Liang, J., 2017. A deep reinforcement learning framework for the financial portfolio management problem.
Jumper, J.M., Evans, R., et al., 2021. Highly accurate protein structure prediction with AlphaFold. Nature 596, 583–589.
Kang, H.-J., Lee, S.-G., Park, S.-Y., 2022. Information efficiency in the cryptocurrency market: The efficient-market hypothesis. J. Comput. Inf. Syst. 62 (3), 622–631.
Kong, W., Liaw, C., Mehta, A., Sivakumar, D., 2019. A new dog learns old tricks: RL finds classic optimization algorithms. In: International Conference on Learning Representations. ICLR.
Laffont, J.-J., Maskin, E.S., 1990. The efficient market hypothesis and insider trading on the stock market. J. Polit. Econ. 98 (1), 70–93.
Lamont, O.A., Thaler, R.H., 2003. Can the market add and subtract? Mispricing in tech stock carve-outs. J. Polit. Econ. 111 (2), 227–268.
Le Tran, V., Leirvik, T., 2020. Efficiency in the markets of crypto-currencies. Finance Res. Lett. 35, 101382.
Lee, K., Kim, S.-A., Choi, J., Lee, S.-W., 2018. Deep reinforcement learning in continuous action spaces: a case study in the game of simulated curling. In: International Conference on Machine Learning. ICLR, pp. 2937–2946.
Li, J., Rao, R., Shi, J., 2018. Learning to trade with deep actor critic methods. In: 2018 11th International Symposium on Computational Intelligence and Design, Vol. 02. ISCID, pp. 66–71.
Li, B., Sahoo, D., Hoi, S.C., 2016. OLPS: A toolbox for on-line portfolio selection. J. Mach. Learn. Res. 17 (35), 1–5.
Li, X., Uysal, A.S., Mulvey, J.M., 2022. Multi-period portfolio optimization using model predictive control with mean-variance and risk parity frameworks. European J. Oper. Res. 299 (3), 1158–1176.
Lillicrap, T.P., Hunt, J.J., Pritzel, A., Heess, N., Erez, T., Tassa, Y., Silver, D., Wierstra, D., 2015. Continuous control with deep reinforcement learning.
Ma, Y., Ahmad, F., Liu, M., Wang, Z., 2020. Portfolio optimization in the era of digital financialization using cryptocurrencies. Technol. Forecast. Soc. Change 161, 120265.
Markowitz, H., 1952. Portfolio selection. J. Finance 7 (1), 77–91.
Mazyavkina, N., Sviridov, S., Ivanov, S., Burnaev, E., 2021. Reinforcement learning for combinatorial optimization: A survey. Comput. Oper. Res. 134, 105400.
Merton, R.C., 1973. An intertemporal capital asset pricing model. Econometrica 867–887.
Moody, J., Saffell, M., 2001. Learning to trade via direct reinforcement. IEEE Trans. Neural Netw. 12 (4), 875–889.
Okoroafor, U.C., Leirvik, T., 2022. Time varying market efficiency in the Brent and WTI crude market. Finance Res. Lett. 45, 102191.
Peng, L., Kloeden, P.E., 2021. Time-consistent portfolio optimization. European J. Oper. Res. 288 (1), 183–193.
Pillay, N., Qu, R., 2018. Hyper-Heuristics: Theory and Applications. Springer Nature.
Pun, C.S., 2018. Time-consistent mean-variance portfolio selection with only risky assets. Econ. Model. 75, 281–292.
Pyun, S., 2019. Variance risk in aggregate stock returns and time-varying return predictability. J. Financ. Econ. 132 (1), 150–174.
Radaideh, M.I., Shirvan, K., 2021. Rule-based reinforcement learning methodology to inform evolutionary algorithms for constrained optimization of engineering applications. Knowl.-Based Syst. 217, 106836.
Rahimian, E., Akartunalı, K., Levine, J., 2017. A hybrid integer programming and variable neighbourhood search algorithm to solve nurse rostering problems. European J. Oper. Res. 258 (2), 411–423.
Rogers, L.C.G., Satchell, S.E., 1991. Estimating variance from high, low and closing prices. Ann. Appl. Probab. 1 (4), 504–512.
Schaerf, A., 2002. Local search techniques for constrained portfolio selection problems. Comput. Econ. 20 (3), 177–190.
Schulman, J., Wolski, F., Dhariwal, P., Radford, A., Klimov, O., 2017. Proximal policy optimization algorithms.
Shajalal, M., Hajek, P., Abedin, M.Z., 2023. Product backorder prediction using deep neural network on imbalanced data. Int. J. Prod. Res. 61 (1), 302–319.
Sharpe, W.F., 1994. The sharpe ratio. J. Portfolio Manag. 21 (1), 49–58.
Shaw, D.X., Liu, S., Kopman, L., 2008. Lagrangian relaxation procedure for cardinality-constrained portfolio optimization. Optim. Methods Softw. 23 (3), 411–420.
Shi, S., Li, J., Li, G., Pan, P., Chen, Q., Sun, Q., 2022. GPM: A graph convolutional network based reinforcement learning framework for portfolio management. Neurocomputing 498, 14–27.
Silver, D., Huang, A., et al., 2016. Mastering the game of go with deep neural networks and tree search. Nature 529 (7587), 484–489.
Silver, D., Schrittwieser, J., et al., 2017. Mastering the game of go without human knowledge. Nature 550 (7676), 354–359.
Sutskever, I., Vinyals, O., Le, Q.V., 2014. Sequence to sequence learning with neural networks. In: Proceedings of the 27th International Conference on Neural Information Processing Systems. NIPS, pp. 3104–3112.
Tao, R., Su, C.-W., Xiao, Y., Dai, K., Khalid, F., 2021. Robo advisors, algorithmic trading and investment management: wonders of fourth industrial revolution in financial markets. Technol. Forecast. Soc. Change 163, 120421.
Thaler, R.H., 1999. The end of behavioral finance. Financ. Anal. J. 55 (6), 12–17.
Tsinaslanidis, P., Guijarro, F., Voukelatos, N., 2022. Automatic identification and evaluation of fibonacci retracements: Empirical evidence from three equity markets. Expert Syst. Appl. 187, 115893.
Vinyals, O., Babuschkin, I., et al., 2019. Grandmaster level in StarCraft II using multi-agent reinforcement learning. Nature 575 (7782), 350–354.
Wilder, J., 1978. New Concepts in Technical Trading Systems. Trend Research.
Woodside-Oriakhi, M., Lucas, C., Beasley, J.E., 2011. Heuristic algorithms for the cardinality constrained efficient frontier. European J. Oper. Res. 213 (3), 538–550.
Wu, X., Chen, H., Wang, J., Troiano, L., Loia, V., Fujita, H., 2020. Adaptive stock trading strategies with deep reinforcement learning methods. Inform. Sci. 538, 142–158.
Wu, Q., Liu, X., Qin, J., Zhou, L., Mardani, A., Deveci, M., 2022. An integrated multi-criteria decision-making and multi-objective optimization model for socially responsible portfolio selection. Technol. Forecast. Soc. Change 184, 121977.
Ye, Y., Pei, H., Wang, B., Chen, P., Zhu, Y., Xiao, J., Li, B., 2020. Reinforcement-learning based portfolio management with augmented asset movement prediction states. In: The Thirty-Fourth Conference on Artificial Intelligence. AAAI, pp. 1112–1119.
Zhang, Y., Bai, R., Qu, R., Tu, C., Jin, J., 2021. A deep reinforcement learning based hyper-heuristic for combinatorial optimisation with uncertainties. European J. Oper. Res..

Dr. Tianxiang Cui is an assistant professor in the School of Computer Science at the University of Nottingham Ningbo China (UNNC) and a senior member of IEEE. Before joining UNNC, he was a senior AI engineer in Huawei and a senior algorithm researcher in PingAn. He was involved in some frontier industrial projects, including autonomous driving and quantitative trading. His main research interests include computational intelligence, particularly metaheuristic, evolutionary computation and neural networks; machine learning and reinforcement learning. He has published a number of research papers in high quality academic journals, including Economic Modeling, International Journal of Production Research, International Review of Financial Analysis, Research in International Business and Finance, Resources Policy, Soft Computing, etc.

Nanjiang Du is now a PhD student in computer science at the University of Nottingham Ningbo China (UNNC). His main research interest includes computational finance, machine learning and federal learning.

Xiaoying Yang is now a PhD student in computer science at the University of Nottingham Ningbo China (UNNC). Her main research interest includes computational intelligence, operations research, machine learning and reinforcement learning.

Dr. Shusheng Ding is a Ph.D. in finance and currently work as an assistant professor in finance in Business School at Ningbo University China. He previously worked as a postdoctoral research fellow at University of Nottingham Ningbo China. His research mainly focuses on the financial modeling and trading with machine learning and deep learning, blockchain financing, financial risk management, volatility forecasting and financial econometrics. He has published a number of research papers in high quality academic journals, including British Journal of Management, Soft Computing, Quantitative Finance, Journal of Futures Markets and Economic Modeling.