Research in International Business and Finance 65 (2023) 101936

Contents lists available at ScienceDirect

Research in International Business and Finance

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

Does reinforcement learning outperform deep learning and traditional portfolio optimization models in frontier and developed financial markets?

Vu Minh Ngo [a,*,1], Huan Huu Nguyen [a,2], Phuc Van Nguyen [b,c,3]

[a] University of Economics Ho Chi Minh City (UEH), Business College, School of Banking, 59C Nguyen Dinh Chieu Street, Ward 6, District 3, Ho Chi Minh City, Viet Nam
[b] Massey University (New Zealand), Private Bag 11 222, Palmerston North 4442, New Zealand
[c] Ho Chi Minh University of Banking (HUB), 36 Ton That Dam Street, Nguyen Thai Binh Ward, District 1, Ho Chi Minh City, Viet Nam

ARTICLE INFO

Keywords:
reinforcement learning
machine learning
portfolio construction models
Sharpe ratio
mean-variance models

ABSTRACT

Advancements in machine learning have opened up a wide range of new possibilities for using advanced computer algorithms, such as reinforcement learning in portfolio risk management. However, very little evidence has been provided on the superior performance of reinforcement learning models over traditional optimization models following the mean-variance framework in different financial market settings. This study uses two experiments with data from the Vietnamese and U.S. securities markets to justify whether advanced machine learning models could outperform traditional portfolios' cumulative returns while optimizing the Sharpe ratio. The results suggest that reinforcement learning consistently outperforms the established methods and benchmarks in both experiments, even when using a very similar degree of diversification in portfolio construction and the same input data. This study confirms the ability of reinforcement learning to provide dynamic responses to market conditions and redefine the risk-return standard in the financial system.

1. Introduction

Portfolio optimization is an integral component of any asset management system. The optimizer seeks to select the optimal asset distribution within a portfolio to maximize the returns at a given risk level. This approach was pioneered in Markowitz's fundamental work (Markowitz, 1952) and is commonly regarded as the modern portfolio theory (MPT). The major advantage of developing such a portfolio comes from the encouragement of diversity that smooths the equity curve, giving a larger return per risk than trading an individual asset. This finding proves that the risk (volatility) of a long-only portfolio is always lower than that of an individual asset for a given projected return as long as assets are not fully connected.

Despite the indisputable potential of such diversification, it is rarely uncomplicated to choose the "right" asset allocations in a portfolio since the dynamics of financial markets change dramatically over time. Assets displaying, for example, high negative

* Corresponding author:
E-mail addresses: vunm@ueh.edu.vn (V.M. Ngo), huannguyen@ueh.edu.vn (H.H. Nguyen), p.nguyen@massey.ac.nz (P. Van Nguyen).
[1] orcid.org/0000-0002-0997-4720
[2] orcid.org/0000-0003-1134-8072
[3] orcid.org/0000-0002-7144-7550

https://doi.org/10.1016/j.ribaf.2023.101936
Received 30 December 2021; Received in revised form 1 January 2023; Accepted 17 March 2023
Available online 25 March 2023

correlations in the past might be favorably linked in the future. This adds additional risk to the portfolio and impairs its future performance. In addition, the range of possible assets for portfolio construction is vast. Using the U.S. stock markets as an example, more than 5000 stocks are offered. Indeed, a well-balanced portfolio not only includes stocks but is also often complemented by bonds and commodities, substantially broadening the range of options.

To address the limitations of quadratic optimization used in traditional mean-variance methods, novel approaches have been tested. However, current research has mostly focused on statistical approaches. Statistical approaches include autoregressive conditional heteroscedasticity (ARCH) (Engle, 1982), autoregressive integrated moving average (ARIMA), and generalized autoregressive conditional heteroscedasticity (GARCH) (Bollerslev, 1986). Machine learning methods aiming at forecasting approaches have also recently become a popular choice in portfolio optimization. Neural networks (NNs) (Bisoi et al., 2019), support vector regression (SVR) (Chen et al., 2015), and ensemble learning (Zhou et al., 2019) are frequently used approaches. According to several comparative studies, machine learning is more capable of dealing with non-linear and non-stationary situations than statistical methods (Wang et al., 2020b).

In additon to machine learning models to forecast future returns for portfolio construction, in this study, we propose advanced methods that directly optimize a portfolio using reinforcement learning. Unlike traditional techniques (McNeil et al., 2015), which begin by forecasting projected returns (usually using econometric models), we skip this phase and obtain asset allocations immediately. Numerous studies (Moody et al., 1998) have shown that the return forecasting technique is not guaranteed to maximize portfolio performance because the prediction stages seek to minimize a prediction loss that is not equal to the portfolio's total gain. In contrast, our method directly optimizes the Sharpe ratio, thus optimizing the return on risk per unit.

Recently, machine learning has been increasingly used in finance, which might have an effect on how hedge fund managers interpret the risk-reward ratio in financial markets (Wu et al., 2021). Advanced machine learning algorithms have shown impressive results in video games (Mnih et al., 2015) and board gameplay (Silver et al., 2016). Since the computer program AlphaGo defeated Lee Sedol, the strongest human player of the contemporary Go board game, in 2016, financial traders have developed a special interest in reinforcement learning (Khushi and Terry Lingze, 2019). Machine learning methods can be utilized for portfolio construction, with reinforcement learning believed to play a more crucial role in the industry (Bartram et al., 2021). Regarding wealth management, Li et al. (2020) highlight that reinforcement learning can be employed effectively in asset allocation.

Reinforcement learning has begun to be used in algorithmic trading (Wen et al., 2021). In options trading, Wen et al. (2021) showed that the reinforcement learning model could achieve decent returns compared to the conventional buy-and-hold strategy. Focusing on futures contracts, Zhang et al. (2020) proved that the designed trading strategies applying reinforcement learning outperform the time series momentum strategies, generating positive profits in spite of heavy transaction costs. In addition, a machine learning algorithm can be trained to hedge options under realistic situations using reinforcement learning (Kolm and Ritter, 2019). Reinforcement learning picks up the minimal variance hedge based on the transaction cost function that is given.

In reinforcement learning, an agent engages with its surroundings and discovers an ideal course of action through trial and error by using reward and penalty points for successful actions and errors, respectively. It is applied to issues involving sequential decision-making (Li, 2017). While the deep learning model requires substantially large historical datasets, reinforcement learning is more advanced due to its ability to self-learn and self-adjust based on the environment. In a continuously volatile investment environment, the use of a historical dataset (deep learning) seems to be less attractive than a self-adjustment mechanism (reinforcement learning). If it could be successfully applied to asset management and portfolio constructions, reinforcement learning could substantially redefine the risk and return standards in the financial sector. However, until now, there has been very little empirical evidence about applying reinforcement learning models in asset management.

To bridge this gap, this study aims to compare the performance of ten different portfolio construction methods covering all the approaches mentioned above, including mean-variance approaches (equally weighted portfolio, maximizing Sharpe ratio, minimum variance, maximizing decorrelation), statistical methods (hierarchical risk parity, principal component analysis, Holt's smoothing process), deep learning models using deep neural networks, and reinforcement learning models. Financial asset prices are closely related to their volatility over time. Naturally, the predictability of stable stocks is better than that of relatively noisy stocks. Thus, using more than one timeframe and asset type in this experiment could provide a better overall picture of the performance of different portfolio construction methods. Two experimental designs with different timeframes and asset types were developed and used in this study to evaluate the performance of these portfolio construction methods in different contexts. In addition, to improve the comparability between methods, historical close prices of financial assets are the only input for all models to control for the ability to process and combine different inputs in non-linear relationships (deep learning and reinforcement learning) to produce much better results than other linear methods.

Moreover, frontier financial markets like the Vietnamese stock exchange have many distinctive features and are much more unstable than developed financial markets like the New York Stock Exchange (NYSE). While the U.S. stock market is the largest and arguably the most developed in the world (Statista, 2022; Wang et al., 2021), Vietnam's stock market is frontier and fast-growing. Initially, only two companies were listed in the Vietnamese stock market; by 2020, this had increased dramatically to 745 domestic listed companies on the two major stock exchanges – the Ho Chi Minh City Stock Exchange (HOSE) and the Ha Noi Stock Exchange (HNX). In 2020, the International Monetary Fund (IMF) predicted that Vietnam's economy would be one of the fastest-growing in Asia with its capacity to avoid a potential recession that might afflict many Asian countries (Giang and Yap, 2020).

Apart from expanding quickly, the Vietnamese stock market is also increasingly drawing investor interest. The Vietnamese government has concentrated on promoting the country's stock market since 2014 in order to attract more foreign investment (Anh and Christopher, 2020). For example, the maximum foreign ownership limitations for publicly traded non-financial firms were previously set at 49%. Since 2016, non-restricted companies' foreign ownership ceiling has been lifted to 100% (Vo et al., 2018). In addition,

Vietnam has consistently strengthened its legal system, provided new securities products, and enhanced the market environment (Anh and Christopher, 2020). As a result, at the end of 2019, the market capitalization of the equity market was around USD 190 billion (79% of GDP), an increase of nearly four times when compared to USD 49 billion in 2014 (31% of GDP) (Anh and Christopher, 2020).

As reported by Bloomberg, the Vietnamese stock market was the top performer in the international and Asian markets throughout the COVID-19 pandemic (Giang, 2020; Giang and Yap, 2020). The market's recovery is being positively fueled by domestic financial flow, particularly from new investors entering the market, according to the State Securities Commission of Vietnam (The Ministry of Finance, 2020). In particular, individual investors registered 31,418 new securities accounts in September 2020, bringing the total for the three quarters of 2020–252,026 accounts, a 34% increase over the number of newly registered accounts in 2019. This resulted in the trading volume on the Vietnamese stock market increased by approximately 40% in the last two quarters of 2020 compared to the same period the previous year (The Ministry of Finance, 2020).

With the huge number of new individual investors entering the market, these inadequate financial literacy investors would benefit from the new approach of managing risks and optimizing portfolio performance, like reinforcement learning and deep learning models. In particular, reinforcement learning enables the "prediction" and "portfolio design" tasks to be combined into a single integrated stage, facilitating investors in achieving their goals. Concomitantly, important constraints (the investor's level of risk aversion, market liquidity, transaction costs) can also be conveniently taken into consideration. Additionally, for other frontier emerging markets in the CIVETS group (Colombia, Indonesia, Vietnam, Egypt, Turkey, and South Africa), the Vietnamese case study of the verification of the reinforcement learning approach might serve as a great example.

Providing two experiments using a variety of optimization methods and in different market conditions, this study demonstrates that the reinforcement learning model has the highest portfolio return and Sharpe ratio performance. More importantly, it consistently outperforms conventional approaches independent of market circumstances providing by the two experiments on Vietnamese stock markets and U.S. stock market. The deep learning models, on the other hand, perform rather well when market conditions are favorable for trading but lose their advantage when market conditions are adverse. This discovery raises the issue of whether deep learning models such as long short-term memory (LSTM) may be superior to conventional approaches, given that deep learning is sensitive to data distribution and market conditions. Second, the variations in performance amongst the models in this investigation may be attributable to the composition of the portfolios. When constructing a portfolio, models like reinforcement learning, Holt's smoothing, and conventional mean-variance use a method known as "high diversification." Other approaches, including the deep learning model, hierarchical risk parity (HRP), principal component analysis (PCA), and max decorrelation, concentrate on selecting a limited number of assets and giving them the greatest portfolio weight. Intriguingly, reinforcement learning models consistently outperform classic mean-variance models in terms of the Sharpe ratio, Sortino ratio, and cumulative returns, even though they both optimize portfolio diversity using the same technique.

The remainder of this paper is organized as follows. Section 2 introduces the pertinent literature and Section 3 details our techniques and data. Section 4 covers our experiments and compares deep learning models and reinforcement learning against traditional optimization and statistical methods. Section 5 discusses key findings, summarizes the results and proposes the potential for future research.

2. Literature review

2.1. Baseline method and optimization portfolio construction

Naïve diversification (equal-weighted portfolio), an example of a baseline portfolio construction approach, is a simple but powerful approach for effectively reducing a portfolio's idiosyncratic risk without harming the expected rate of return. Different from naïve diversification, optimization measures, pioneered by the Markowitz theory, refer to an optimal set of weights as one in which the portfolio achieves an acceptable baseline expected rate of return with minimal volatility (Markowitz, 1952).

The modern portfolio theory (MPT) developed by Markowitz (1952) states that returns are maximized for a given level of risk based on mean-variance portfolio construction. The efficient frontier is the foundation of this theory, which mentions how investors create a portfolio that maximizes expected returns based on a specific level of risk. Markowitz's efficient frontier is the portion of the minimum-variance curve that lies above and to the right of the global minimum variance portfolio, preferred by rational, risk-averse investors. When the risk level increases, the increasing curve begins to flatten. This means that we cannot achieve ever-increasing returns as we take on more risks.

It is widely noted that market capitalization-weighted portfolios are inefficient and underperform an equally weighted portfolio over the long term (Bolognesi et al., 2013; Malladi and Fabozzi, 2017). Furthermore, Demiguel et al. (2009) showed that not only is the equal-weighted portfolio more efficient than the capitalization-weighted portfolio, but it also outperforms mean-variance-based portfolio strategies out of sample.

However, recent studies show that an equal-weighted portfolio of stocks in the S&P500 significantly underperforms the market capitalization-weighted portfolio (Taljaard and Maré, 2021). In particular, an equal-weighted portfolio underperforms the market capitalization-weighted portfolio in the short term. Additionally, Kritzman et al. (2010) argued that optimized portfolios generate better out-of-sample performance compared to equal-weighted portfolios. Therefore, we consider all of these stock construction approaches, including the baseline and optimization methods, to provide a comprehensive comparison.

2.2. Applying machine learning in finance

The Markowitz efficient frontier is built on assumptions that can be criticized in realistic situations (Ma, 2021). For example, it assumes that all investors are rational and risk-averse and that all investors are equally able to borrow money at a risk-free interest rate, although this is not the case in reality. Additionally, a proto-idea of an efficient frontier is that asset returns follow a normal distribution, whereas, in reality, asset returns often vary three standard deviations away from the mean.

Data mining is the search for replicable patterns, typically for large datasets, from which we can derive benefits that are increasingly important in finance (Arnott et al., 2019). In the past, researchers had to pay substantial attention to the relevant hypotheses that made the most sense. In recent years, in the era of machine learning, researchers have not specified a hypothesis that will be potentially determined by the algorithm. Financial data are particularly suited for deep learning because of the substantial amount of data available for training (Soleymani and Paquet, 2020). Recently, applications of deep learning to predict financial market data have achieved significant success (Chong et al., 2017; Lahmiri and Bekiros, 2020; Long et al., 2019).

Machine learning holds considerable promise for the development of successful trading strategies, especially in higher-frequency trading, in addition to being impractical to use in the past (Arnott et al., 2019). The advantages of applying algorithmic machine learning in trading are widespread (Zhang et al., 2020a) for finding higher alpha (excess returns) (Sirignano and Cont, 2019; Zhang et al., 2019). Most research focuses on regression and classification pipelines, in which excess returns or market movements are predicted over certain (fixed) horizons.

2.3. Machine learning and portfolio risk management

Machine learning has been increasingly applied in finance, potentially affecting the way hedge fund managers define the risk-reward ratio in the financial market. Given their flexibility, hedge funds have attracted increasing investor attention (Wu et al., 2021). In a comprehensive study, Wu et al. (2021) applied machine learning to hedge fund return prediction and selection and showed that machine learning-based forecast methods outperform the respective styled Hedge Fund Research indices in almost all situations. Among the four machine learning methods in Wu et al. (2021), neural networks stand out in general.

Regarding risk management, Arroyo et al. (2019) shows that machine learning helps venture capital investors in their decision-making processes to find opportunities and assess the risk of investments. In addition, Jurczenko (2020) found that machine learning algorithms are useful in improving stock risk forecasts, especially out-of-sample equity beta predictions.

In a comprehensive study by Gu et al. (2020), machine learning tools outperformed linear methods in terms of their predictive capability. The efficiency of portfolios built using machine learning algorithms has been clarified for unoptimized portfolios (Kaczmarek and Perez, 2021). Despite its unquestionable popularity, the modern portfolio theory has been criticized for its unreliability in practice (Kolm et al., 2014; DeMiguel et al., 2009). Therefore, an emerging branch of literature focuses on portfolio optimization enhancements, including replacing the statistical moments of asset returns with a more reliable prediction (DeMiguel et al., 2009) or modern mathematics by applying machine learning instead of quadratic optimization (De Prado, 2016).

Regarding optimized approaches, Kaczmarek, Perez (2021) prove that when the stocks of portfolios are preselected by machine learning methods, traditional portfolio optimization techniques (mean-variance and hierarchical risk parity) also increase the risk-adjusted return of these portfolios, outperforming the equal-weighted portfolios out of the sample.

Recently, reinforcement learning methods have received attention for portfolio allocation tasks. Since the seminal work of Sutton and Barto (2018), there has been an unprecedented emerging branch of finance on the application of reinforcement learning methods to portfolio construction. It was initially applied to cryptocurrencies, the Chinese financial market (Liang et al., 2018; Yu et al., 2019; Wang & Zhou, 2020; Saltiel et al., 2020), and other assets (Kolm and Ritter, 2019; Liu et al., 2020; Ye et al., 2020; Li et al., 2016; Xiong et al., 2019).

3. Experimental design

3.1. Portfolio construction methods

3.1.1. Mean-variance method

3.1.1.1. Maximizing sharpe ratio portfolio (Max_Sharpe ratio). On the traditional methods side, the seminal work of Markowitz (1952) has led to various extensions, such as minimum variance (Chopra and Ziemba, 1993; Kritzman et al., 2010), maximum diversification (Choueifaty et al., 2012), and maximum decorrelation (Christoffersen et al., 2010). Traditional methods are based on maximizing the Sharpe ratio, which is defined as.

Sharpe ratio $= \frac{R_p - R_f}{\sigma_p}$ where

R_p : *returnofportfolio*

R_f : *riskfreerate*

σ_p : *standarddeviationoftheportfolio*$'$ *esscessreturn*

The Sharpe ratio was developed based on modern portfolio theory (Elton and Gruber, 1997). Adding diversification can increase the Sharpe ratio, resulting in a higher return with the same risk, or a lower risk with the same return level (Sharpe, 1994). The Sharpe ratio explains whether a portfolio's excess returns are due to smart investment decisions or the result of excessive risk. Although one portfolio or fund can enjoy higher returns than its peers, it will be a good investment if those higher returns do not come with excess additional risk.

We try to find the maximum Sharpe ratio for a given minimum return r_{min} as follows:

$$Minimize_w w^T \sum w$$

subject to$\mu^T w >= r_{\min}, \sum_{i=1}^{n} w_i = 1, 1 \geq w \geq 0$where w = (w1; …; wl) with 1 ≥ wi ≥0 for i = 1 …n, is the allocation weight for each asset in the portfolio and all the weights sum to 1: $\sum_{i=1}^{l} w_i = 1$. μ = (μ1,., μL)T is the expected returns vector and $\sum$ is the variance-covariance matrix of n assets in the portfolio.

3.1.1.2. Minimum variance portfolio (Min_Var). A minimum variance portfolio suggests building a portfolio with a list of securities with minimum price volatility. The term originates from the Markowitz portfolio theory (Markowitz, 1991), which implies that volatility can be used to replace risk and, therefore, less volatility variance correlates with less investment risk. The optimization of the minimum variance portfolio is as follows:

$$Minimize_w w^T \sum w$$

subject to $\sum_{i=1}^{n} w_i = 1, 1 \geq w \geq 0$where w_i is the weight allocated to stock i in the portfolio of n stocks, $\sum$ is the variance-covariance matrix of assets in the portfolio.

3.1.1.3. Maximum decorrelation portfolio (Max_Decorr). The maximum decorrelation portfolio shows that an investor believes that all assets have similar returns and volatility but heterogeneous correlations (Christoffersen et al., 2010). This theory aims to optimize the minimum variance, which is performed on the correlation matrix instead of the covariance matrix. The weights obtained from the maximum correlation optimization are divided by the respective degrees of flight and renormalized so that they sum to 1 to optimize the weighting of the portfolio. If we denote ℂ as the correlation matrix of portfolio assets, the optimization program for the maximum decorrelation portfolio is as follows:

$$Minimize_w w^T \mathbb{C} w$$

subject to $\sum_{i=1}^{n} w_i = 1, 1 \geq w \geq 0$where w_i is the weight allocated to stock i in the portfolio of n stocks, ℂ is the correlation matrix of n assets in the portfolio.

3.1.2. Deep learning portfolio (deep portfolio)

Some popular deep learning models include the fully connected neural network (FCN), the convolutional neural network (CNN), and long short-term memory (LSTM) based on the recurrent neuron network (RNN) (see Appendix 1 for more discussion on RNN) (Sepp and Jürgen, 1997). In general, LSTMs perform best when modeling daily financial data because of their design nature for processing and predicting sequential data (Zhang et al., 2020b). Thus, a deep learning model based on LSTM was employed to directly optimize portfolio performance.

Financial time series used in this study may be seen as the following three-dimensional tensor including time, asset and indicators. For example, one may examine numerous stocks' (asset dimension) daily (time dimension) close price returns, closing price returns, and volumes (channel dimension) (asset dimension). We divided the three-dimensional tensor into three parts:

1. Input x: all historical and contemporary knowledge of the financial time series and used as input for the weight allocation prediction.
2. Gap g: reflects immediate future knowledge of the financial time series that cannot be used to make investment choices.
3. Future horizon y: the development of the market in the future which is used to calculate the return of the portfolio with allocation weight w.

This study employs 30 assets for each portfolio. The lookback period is 250 trading days and the gap is 5 days. Close price and daily returns calculated from the close price are indicators in the sample. This study examines deep learning networks that receive an input x (the three-dimensional tensor mentioned above) and produce a single weight allocation w of shape n, where n is the number of assets in x. Then, assuming F is a neural network with parameters θ, the prediction process could be generally described as:

$$F(x, \theta) = w$$

where the allocation weight for stock i is $w_{i,t} \in [0, 1]$ and $\sum_i^n w_{i,t} = 1$.

The objective function inputs w and y to produce an output as a real number. This study used the Sharpe ratio as the algorithm for calculating the objective function for the deep learning models. The loss function L could be generally described as:

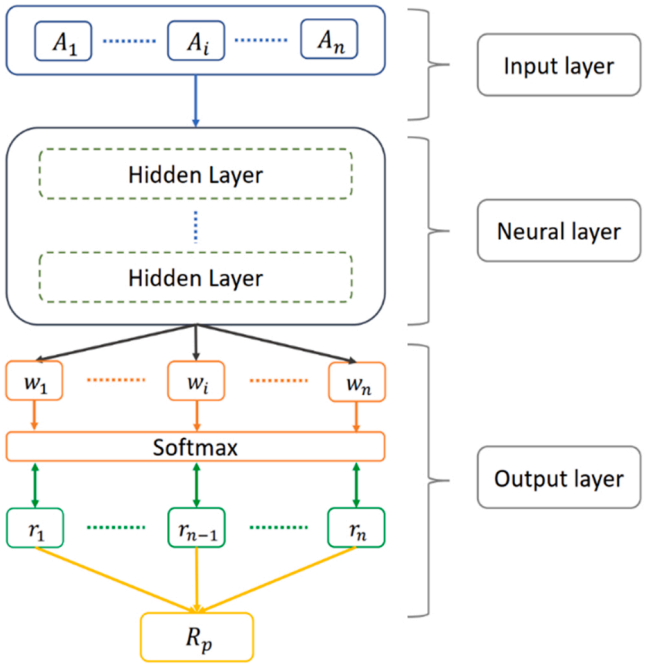

Fig. 1. : Deep learning model architecture.
Source: Zhang et al. (2020b).

$L(w,y) = L(F(x,\theta),y) = \frac{E(R_p)}{Std(R_p)}$where $E(R_p)$ and $Std(R_p)$ are estimates of the portfolio's mean and standard deviation, respectively. The gradient ascent is used in this study to maximize the objective function of this Sharpe ratio. When we have the gradient of the Sharpe ratio according to the set of parameters θ: $\frac{\partial L}{\partial \theta}$, this value could be repeatedly calculated from training data and could update the parameters using gradient ascent: $\theta_{new} := \theta_{old} + \alpha \frac{\partial L}{\partial \theta}$, where α is the learning rate and the process can be repeated for a large number of epochs until convergence of the Sharpe ratio.

Specifically, the process includes three main blocks: the input, neural, and output layers. First, we concatenate the features (past prices and returns) of all assets and then feed this input into the network. The input layers consist of a three-dimensional tensor ($A_1 \ldots A_j \ldots A_n$) as depicted in Fig. 1.

Second, a neural network, which is a series of stacked hidden layers, is used to extract cross-sectional features from the input. Each hidden layer consists of multiple LSTM units (Fig. A2) in each timestep (see Appendix 1 and 2 for more discussion on the recurrent neuron network (RNN) and LSTM models and other deep learning models).

Finally, the model outputs the number of nodes equal to that of the assets in our portfolio, which are portfolio weights. These weights can be multiplied by the associated assets' returns to gain realized portfolio returns and maximize the Sharpe ratio. In addition, because of using a long-only portfolio, we use softmax functions to achieve this requirement:

$$r_{i,t} = \frac{EXP(\widetilde{w_{i,t}})}{\sum_{j}^{n} EXP(\widetilde{w_{j,t}})}$$

where $(\widetilde{w_{i,t}})$ is the raw weight of stock i in the portfolio and $r_{i,t}$ is the final weight for the long-only porfolio using the softmax transformation (Fig. 1).

3.1.3. Reinforcement learning

Typically, value-based and policy-based algorithms are used to classify solutions to the reinforcement learning issue. Two well-known algorithms are often used to elaborate and illustrate the fundamentals of reinforcement learning. These are the Q-learning and SARSA algorithms. Given its popularity and effectiveness in many applications, this study also employed the Q-learning function for the reinforcement learning portfolio.

In its updating action and state for the future, Q-learning assumes the optimum policy. The usage of the $\max_a$function over the available actions renders the Q-learning algorithm non-compliant with the policy. This is because the policy we are updating behaves differently from the policy we use to explore the environment, which utilizes an exploration parameter called epsilon to pick between the best-identified action and a random action:

$$Q(s_t, a_t) \leftarrow Q(s_t, a_t) + \alpha[R_t + \gamma \max_a Q(s_{t+1}, a) - Q(s_t, a_t)](*)$$

where R_t is the reward received when taking action a_t at state s_t, α is the learning rate ($0 < \alpha < 1$), and γ is the discount coefficient measuring the value placed on future benefits ($0 < \gamma < 1$). If γ is closer to zero, the agent is more likely to prioritize instant benefits. If γ is closer to one, the agent will give higher weight to future benefits and be ready to defer the present reward.

The steps for Q-learning are as follows:

1. Step 1: At time t, start from the state s_t and pick an action a_t.We apply an ε *−greedy* approach to randomly select an action with a probability of ε; otherwise, choose the best action to maximize the Q-value function, $a_t = \max_a Q(s_t, a)$.
2. Step 2: With action a_t, we get reward R_t and get to the next state s_{t+1}.
3. Step 3: At state s_{t+1}, choose the best action and get the value of the Q-function as $\max_a Q(s_{t+1}, a)$.
4. Step 4: We update the Q-learning function (*).
5. Step 5: Repeat the steps to optimize the course of action at different states.

As is typical in reinforcement learning, we modeled our algorithm to be solved using a Markov decision process (MDP), as described by Sutton and Barto (1998). The MDP presupposes that the agent is aware of all possible states of the environment and has all the necessary knowledge to make the optimum choice in each condition. Additionally, the Markov property indicates that knowledge of the present state is sufficient.

MDP assumes a four-tuple (S, A, P, R), where S is the set of states, A is the set of actions, P is the probability function $P : S_t x\ A\ x S_{t+1} -> [0; 1]$from state action ($S_t x\ A$) to the next state transition S_{t+1}, and R is the immediate reward. The agent's objective is to learn a policy (π :courses of action) that maps states to the best action $\pi : S \rightarrow A$ while optimizing the anticipated discounted Sharpe ratio.

Reinforcement learning is applied to portfolio management as follows:

1. The current market knowledge is codified using a state variable represented by s_t.
2. For the weight allocation in the portfolio construction problem, state s_t is the correlation matrix of the instruments based on a lookback period at time t.
3. Our planning goal, which is to determine the best portfolio allocation, may be thought of as engaging in market activity through action a_t. This action corresponds to the current portfolio allocation choice (also known as allocation weights).
4. After the portfolio allocation has been determined, we see the following state s_{t+1}.
5. We utilize the daily return to assess the effectiveness of our action a_t. In this case, we can calculate the reward for an action. In this study, the Sharpe ratio was used as the reward for the reinforcement learning algorithm. Reward R_t is the Sharpe ratio of the portfolio, given the portfolio weight and lookback period.
6. We update the Q-learning function in (*) and repeat the process to optimize the weight allocation for different covariance matrices of assets in the portfolio.

In comparison to earlier conventional portfolio allocation approaches, the fundamental optimization issue in reinforcement learning is distinct: (1) First, we optimize a policy π, not basic weights w_i for asset i in the portfolio. Although this function is ultimately represented by a deep neural network with weights, the approach is fundamentally different because we are optimizing in the space of functions:$\pi : S \rightarrow A$, which is a considerably larger space than merely the return R. (2) Second, it is a multi-timestep optimization because it incorporates findings from time $t = 1$ to time $t = T$ (termination value or the end of the training data), which makes it more engaging in the dynamics of the trading environment.

3.1.4. Other statistical methods

3.1.4.1. Hierarchical risk parity (HRP). The HRP algorithm seeks to address some of these issues using a novel technique based on the idea of a hierarchy (De Prado, 2016). This algorithm comprises three distinct stages:

- Hierarchical clustering - categorizing our stuff hierarchically
- Quasi-diagonalization - restructuring the covariance matrix to group together assets with comparable attributes
- Recursive split - giving weights to each asset in our portfolio.

This contributes to minimizing the limits of Markowitz's MPT theory, allowing it to be implemented in reality, such as predicting the return on a specific set of assets or computing the inverse of the covariance matrix, which affects the algorithms and, hence, the model outcomes.

3.1.4.2. Principal component analysis (PCA). Principal component analysis (PCA) is a frequently used technique for dimensionality reduction (Partovi and Caputo, 2004). We identify the structure of the covariance or correlation matrix and utilize it to establish low-dimensional subspaces containing the majority of portfolio fluctuations. This is the first time the concept of employing PCA for efficient portfolio analysis has been proposed. The fundamental premise of this theory is that if there are no connections between

Table 1
Summary statistics for close prices of assets used.

Stocks	Mean	Std. Dev.	Min	Max	ETFs	Mean	Std. Dev.	Min	Max
BID	39.96589	4.878451	29.59	54.58	AGG	110.0235	1.304483	107.36	113.25
CTD	73.15809	17.89981	42.6152	134.2285	DIA	174.1366	6.883296	156.49	186.52
CTG	22.64907	7.720064	13.1382	42.1477	EEM	37.97426	4.293147	28.25	45.85
EIB	19.45136	4.037072	14.65	33.4	EFA	61.88778	4.433597	51.38	70.67
FPT	51.38163	21.04408	27.6193	99.4	EWT	14.85896	1.881408	11.34	31.52
GAS	84.34686	13.38598	49.8083	125	EWU	35.71888	4.111879	27.98	44.04
HDB	16.87453	5.537896	8.3638	29.92	EWY	55.94181	5.472162	44.71	67.36
HPG	25.03047	14.54703	9.7447	58	EWZ	34.06847	9.302373	17.33	54
MBB	16.75488	6.495295	8.6483	32.1852	GLD	117.1061	7.394246	100.5	130.52
MSN	81.94171	26.24323	48.0061	151.7	IAU	11.80742	0.7458444	10.15	13.18
MWG	76.9147	21.21408	38.6061	133.9	IVV	204.8055	8.793418	182.6	220.37
NVL	57.41127	24.65236	37.1648	121	IWD	100.9484	4.036954	87.41	107.33
PLX	50.88562	5.311053	32.9815	60.2153	IWM	116.168	6.409031	94.79	129.01
PNJ	77.0555	13.62735	45.1111	106.5	IYR	75.46781	4.124172	66.27	85.7
POW	11.55301	1.844234	6.7632	15.3579	LQD	118.6493	2.746482	112.92	124.4
REE	42.05047	12.25542	27.3	76	OIL	12.11136	7.044019	4.46	25.91
ROS	11.66419	11.00721	2.09	34.8	SLV	16.29648	1.918915	13.06	20.57
SAB	195.7162	39.48405	113.3927	276.9403	SPY	203.5199	8.7182	181.51	219.09
SBT	17.8494	2.953768	11.5254	25	VGK	52.81503	4.443157	43.32	61.72
SSI	17.59558	10.60019	6.2644	44.25	VTI	104.7933	4.495938	92.56	112.75
STB	15.73822	6.971162	7.3	33.8	VWO	38.48137	4.275838	28.55	46.49
TCB	29.71024	12.0473	14.9	58	XLB	47.12634	3.018081	37.28	52.09
VCB	86.01992	12.24443	56.7348	116.4	XLE	75.6521	12.44283	51.77	101.29
VHM	67.46873	9.654534	41.5624	91.0125	XLF	23.23109	1.373564	19.04	25.58
VIC	96.29151	10.51808	63.5556	128	XLI	54.70775	2.346758	48.01	59.08
VJC	122.274	12.95467	94.5	148.2	XLK	41.86613	2.835571	35.2	47.99
VNM	93.53449	8.133067	65.2065	112.922	XLP	49.18243	3.271228	42.64	55.71
VPB	20.15967	10.65923	10.4533	44.835	XLU	45.33947	3.044848	40.17	52.98
VRE	30.70347	3.87047	17.7	37.8	XLV	68.81307	5.124752	55.71	77.22
					XLY	74.21737	5.488598	62.24	82.16

Note: Summary statistics of the closing prices of assets used for the two experiments.

assets, portfolio selection is much simpler. Based on significant portfolios, the theory constructs an efficient frontier. According to a theoretical interpretation of such a shift, the PCA contends that the efficient frontier's return–volatility structure is more directly tied to the main portfolio environment than to the initial asset collection.

3.1.4.3. Holt's smoothing process (Holt's smoothing). Another approach to portfolio construction is to use forecasting methods to produce predicted asset returns from historical time series; then, the weight for each asset is optimized for portfolio allocation. In this research, Holt's exponential smoothing process is chosen as the forecasting method, given the short timeframe of the forecasting sample. Specifically, double exponential smoothing with a linear additive trend (Holt, 1957) was used as follows:

$$s_t = \alpha x_t + (1-\alpha)(s_{t-1} + b_{t-1})$$

$$b_t = \beta(s_t - s_{t-1}) + (1-\beta)(b_{t-1})$$

where s_t is the current value of the series at time t, b_t is the trend component of the time series, and $\alpha(0 < \alpha < 1)$ is the factor for smoothing the time-series data x_t. The larger α is, the larger the impact of current value x_t on the time series value s_t. β is the smoothing factor of the trend. The future data of the time series x_t are forecast using:

$$F_{t+m} = s_t + m.b_t$$

where m is the number of timesteps to forecast and F_{t+m} is the forecasted value of time series in the m timestep in the future.

3.2. Data

In both experiments, the only input data is the historically adjusted close price of assets collected from trading exchanges. By their natural settings, the mean-variance method and its extensions, such as minimum variance, maximum decorrelation, and risk parity, require only a close price of assets as input. By contrast, the deep learning (deep portfolio) and reinforcement learning methods can include many different types of data as inputs in their models. Using close price as the only input in all portfolio construction methods also enables a better comparison of their performance by controlling the input data specifications and quality required by these

Table 2
Portfolio evaluation metrics for VN30 experiment.

	Reinf. learning	Deep portfolio	PCA	HRP	Holt's smoothing	Max_Decorr	Max_ Sharpe ratio	1/N	Min_ Var	Market cap
Sharpe ratio	3.7928	3.3591	1.2703	3.3082	1.6911	2.2982	1.8573	1.8844	1.8566	1.5812
Sortino ratio	5.3098	5.2693	1.9216	5.5661	2.3301	3.5793	2.5383	2.5688	2.5924	2.1442
E(R)	0.0032	0.0024	0.0017	0.0028	0.0011	0.0016	0.0012	0.0012	0.0014	0.0010
Std(R)	0.0294	0.0269	0.0399	0.0296	0.0264	0.0265	0.0263	0.0260	0.0283	0.0263
Max DD	-0.1280	-0.1067	-0.3158	-0.1375	-0.1060	-0.1302	-0.1206	-0.1144	-0.1087	-0.1153
Var (1%)	-0.0193	-0.0164	-0.0395	-0.0200	-0.0171	-0.0167	-0.0168	-0.0164	-0.0195	-0.0170
Range	0.1069	0.0646	0.1144	0.0919	0.0752	0.0639	0.0667	0.0663	0.0835	0.0697
Maximum	0.0412	0.0298	0.0550	0.0439	0.0328	0.0322	0.0263	0.0294	0.0361	0.0318
Minimum	-0.0657	-0.0347	-0.0594	-0.0480	-0.0424	-0.0316	-0.0404	-0.0369	-0.0475	-0.0378
Skewness	5.5474	0.5265	-0.0154	2.1757	2.2060	0.4715	2.1023	1.8781	2.1021	2.1271
Kurtosis	-1.4859	-0.3997	-0.0490	0.0239	-0.7936	-0.1664	-0.9848	-0.9566	-0.8168	-0.9187

Note: This table shows the evaluation metrics for the portfolios constructed from VN30 stocks by deep learning methods (deep portfolio and reinforcement learning), optimization methods (HRP, PCA, and Holt's smoothing), and baseline methods (1/N, Max_Sharpe ratio, Min_Var, Max_Decorr and Market cap weighted). Regarding the evaluation metric, E(R) is mean daily return while Std(R), Max DD, Var (1%) represent for daily risk metrics. While Sortino ratio assesses the downside risk of an investment's return, Sharpe ratio calculates returns based on total market volatility, which takes both upside and downside risks into account. The other statistical measures are used to describe and summarize data. In specific, range is the difference between the largest value (maximum) and smallest value (minimum) in a set of data. While Skewness measures the asymmetry of a distribution, Kurtosis shows the flatness or peakedness of a distribution.

methods.

The extant literature shows that diversification, eliminating investment risk without sacrificing the expected rate of return, provides a benefit until the portfolio size ranges from 10 to 15 stocks.[4] Early on, Sharpe (1970) noted that risk can be significantly decreased by diversifying investments even a little. He proposed that a portfolio with around fifteen securities may be regarded as well-diversified. This conclusion has been reached by Elton and Gruber (1977) from an analytical solution of the link between portfolio size and risk. Similarly, Mokkelbost (1971) also discovered that when "quite few" different securities are included in the portfolio, a significant amount of the attainable reduction is achieved. Campbell et al. (2001) show that a portfolio that has 20 or 30 stocks closely resembles one that is well-diversified and has nearly no unsystematic risk, analogous to the recommendation from Statman (1987). Tang (2004) empirically proved that with an infinite population of stocks, a portfolio construction approach with a size of 20 can eliminate 95% of the diversifiable risk. Moreover, an additional 80 stocks are required to reduce idiosyncratic risk by an extra 4%. Therefore, a portfolio of 30 individual stocks is selected in our research so that investors have sufficient assets for selection to eliminate most idiosyncratic (non-systematic) risks, leaving only systematic risk.

In the first experiment, we used common trading stocks from the Ho Chi Minh Stock Exchange (HOSE). Specifically, we use stocks that comprise the VN30 index (VN30) at the time of data collection. The VN30 index measures the performance of the 30 biggest stocks on HOSE based on their market capitalization and liquidity. The list is evaluated every six months to ensure that all stocks in the VN30 meet the requirements of being the largest stocks on the HOSE, as well as other requirements such as being listed on the HOSE for at least six months, being excluded from the bourse's containment list, and having 10% of its total shares as free-float shares. This experiment selects 29 stocks from the VN30 index as candidate assets (one stock is excluded), a sample size sufficiently large for individual investors to consider before forming portfolios. BVH, the largest state insurance company, was excluded from the list of VN30 during the time of data collection by failing to satisfy requirement of at least 10% free-floating share and failing to achieve unqualified opinion from auditors in 2020. The ticker symbols for the 29 stocks are BID, CTD, CTG, EIB, FPT, GAS, HDB, HPG, MBB, MSN, MWG, NVL, PLX, PNJ, POW, REE, ROS, SAB, SBT, SSI, STB, TCB, VCB, VHM, VIC, VJC, VNM, VPB, and VRE. The list is composed of blue-chip stocks with high turnover and total market value, as well as those with a large scale, relatively stable price structure, and adequate liquidity. Appendix 5 provides correlations map between stocks in the VN30 portfolio.

In the second experiment, exchange trade fund securities (ETF) are used to form a portfolio instead of individual stocks. An ETF is a form of security that follows an index, sector, commodity, or other asset class, but may be acquired or sold on a stock exchange market. The ETF differs from individual stocks in its risk-return setting because it presents the weighted return and volatility of several stocks in an exchange. Thus, it can provide another distinct context for testing. Specifically, we selected 30 ETFs from the NYSE Arca using stratified random sampling. The ticker symbols for the 30 ETFs are AGG, DIA, EEM, EFA, EWT, EWU, EWY, EWZ, GLD, IAU, IVV, IWD, IWM, IYR, LQD, OIL, SLV, SPY, VGK, VTI, VWO, XLB, XLE, XLF, XLI, XLK, XLP, XLU, XLV, and XLY. These 30 ETFs are selected from the pools of more than 100 market-capital leading ETFs of the NYSE Arca to ensure the exclusion of irrational trading behaviors from illiquid ETFs. Moreover, we use the stratified random sampling technique to select ETFs from different sectors to cover a wide range of economic activities and financial assets such as gold, oil, U.S. domestic sectors, developed ex-U.S. markets, emerging markets, broad market ETFs, etc. This practice ensures that the correlations between ETFs are not systematically too high or too low, which potentially

[4] A list of recommendations about the appropriate number of stocks can be retrieved from Table 1 of Tang (2004).

Table 3
Portfolio evaluation metrics for ETF portfolio experiment.

	Reinf. Learning	Deep portfolio	PCA	HRP	Holt's Smoothing	Max_Decorr	Max_ Sharpe ratio	1/N	Min_Var	Market cap
Sharpe ratio	1.3457	-0.1858	1.1862	1.3149	1.5392	-0.0277	0.9368	0.8384	0.3029	0.4151
Sortino ratio	2.1299	-0.2511	1.8332	1.9711	2.6059	-0.0311	1.4922	1.2093	0.4039	0.5632
E(R)	0.0007	-0.0001	0.0009	0.0002	0.0009	0.0000	0.0004	0.0004	0.0001	0.0001
Skewness	10.3473	3.5833	0.1538	0.6669	9.3485	69.251	15.9574	4.7621	5.5712	5.1187
Kutoris	0.7865	-0.6170	0.0952	-0.0724	1.2221	-6.6728	1.3956	-0.3342	-1.0648	-0.9323
Std(R)	0.0227	0.0204	0.0292	0.0120	0.0242	0.0198	0.0212	0.0215	0.0210	0.0199
Max DD	-0.0562	-0.0858	-0.0996	-0.0185	-0.0560	-0.0898	-0.0470	-0.0523	-0.0600	-0.0545
Var (1%)	-0.0128	-0.0110	-0.0213	-0.0036	-0.0144	-0.0103	-0.0113	-0.0117	-0.0113	-0.0101
Range	0.1069	0.0646	0.1144	0.0919	0.0752	0.0639	0.0667	0.0663	0.0835	0.0697
Maximum	0.0412	0.0298	0.0550	0.0439	0.0328	0.0322	0.0263	0.0294	0.0361	0.0318
Minimum	-0.0657	-0.0347	-0.0594	-0.0480	-0.0424	-0.0316	-0.0404	-0.0369	-0.0475	-0.0378

Note: This table shows the evaluation metrics for the ETF portfolios constructed from the NYSE Arca by deep learning methods (deep portfolio and reinforcement learning), optimization methods (HRP, PCA, and Holt's smoothing), and baseline methods (1/N, Max_Sharpe ratio, Min_Var, Max_-Decorr and Market cap weighted). Regarding the evaluation metric, E(R) is mean daily return while Std(R), Max DD, Var (1%) represent for daily risk metrics. While Sortino ratio assesses the downside risk of an investment's return, Sharpe ratio calculates returns based on total market volatility, which takes both upside and downside risks into account. The other statistical measures are used to describe and summarize data. In specific, range is the difference between the largest value (maximum) and smallest value (minimum) in a set of data. While Skewness measures the asymmetry of a distribution, Kurtosis shows the flatness or peakedness of a distribution.

underperform and overperform, respectively the portfolio optimization performance. A heatmap of correlation between the 30 ETFs is provided in Appendix 5.

The data collection periods were March 2019 to October 2021 for Experiment 1 and April 2014 to November 2016 for Experiment 2. The duration of the first experiment included bull markets on both the Vietnamese market and golablly. During this time span, the VNINDEX rose from around 990–1450 points. In contrast, the period of the second experiment included markets that are quite steady. During this time, the S&P500 index hovered between 1850 and 2150 points before soaring above 2800 points after November 2016. Using these various periods to account for diverse trend settings in stock markets, we would want to remove the effects of time and particular events on investing results.

To conduct the analysis and performance evaluation, the dataset was divided into training and testing sets. The training set comprises approximately two years of trading days (approximately 500 trading days). It was used to obtain the parameters for all portfolio construction models. The test set comprises about 150 days (six months of trading days) to evaluate and compare the performance between 10 different methods of portfolio construction. Table 1 presents the summary statistics of the closing prices for the two experiments.

4. Experimental results

This study provides two experiments using two distinct market conditions using Vietnamese and American stock markets to test whether reinforcement learning model consistently outperforms other deep learning models and conventional models in terms of portfolio performance. In addition, we further test these models' performances using different timeframe to exclude the time-effect's and specific events' impacts on investment outcomes.

We used the expected return (E(R)), standard deviation of return (Std(R)), Sharpe ratio, Sortino ratio, maximum drawdown (Max DD), and value-at-risk (VAR-1%) to evaluate the performance of different portfolios according to different optimization techniques. In addition, we report some additional metrics for the daily returns for each portfolio, such as range, skewness, and kurtosis, as presented in Tables 2 and 3.

4.1. .Experiment 1: Vietnamese stock market (VN30)

Table 2 presents the evaluation metrics for the portfolios constructed from VN30 stocks. Among the baseline methods (1/N, Max_Sharpe ratio, Min_Var, Max_Decorr and Market cap weighted), optimization methods (HRP, PCA, and Holt's smoothing), and deep learning methods (deep portfolio and reinforcement learning), the reinforcement learning portfolio has the highest daily mean return of 0.0032, followed by the HRP portfolio, with a return of 0.0028, and then the deep portfolio, with 0.0024. Baseline portfolios have mean daily returns from 0.0010 to 0.0016, which are significantly lower than those of the machine learning methods mentioned above.

Regarding the daily risk characteristics, we use the standard deviation of daily stock returns, value at risk (1%), and maximum drawdown to measure the risks of the constructed portfolios. We can see that daily volatility varies in a narrow range from approximately 0.0260 to 0.0296, except for the exceptionally high volatility of the PCA portfolio (0.039). The same pattern was observed for Max DD and var at risk (1%). It can be seen that reinforcement learning and deep portfolios have slightly higher volatility of return

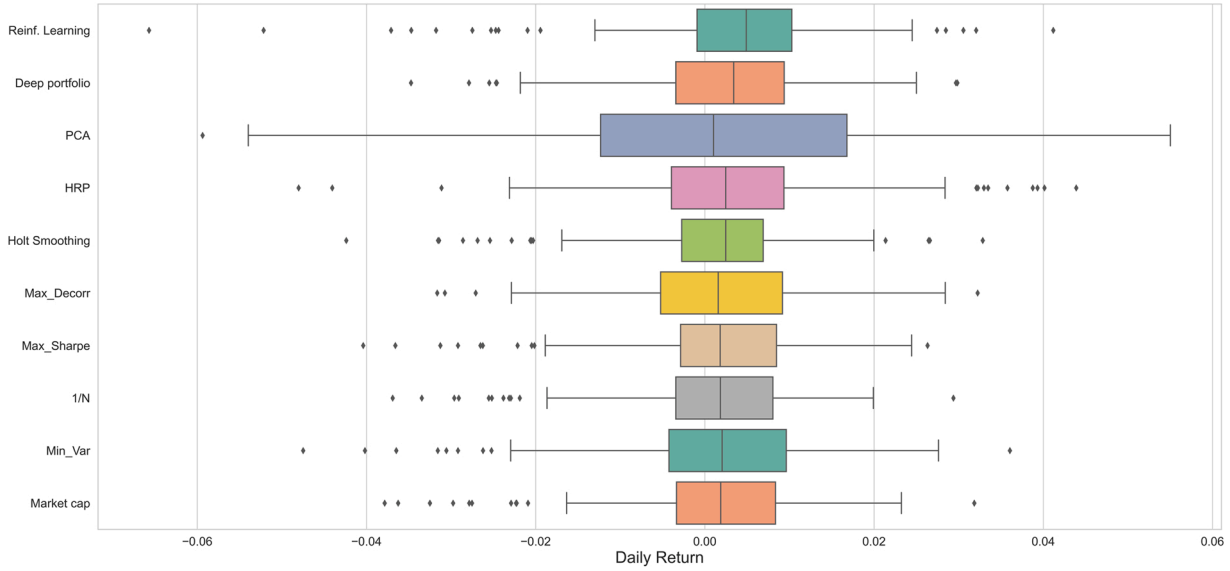

Fig. 2. Box plot of the daily returns of different portfolio construction methods.

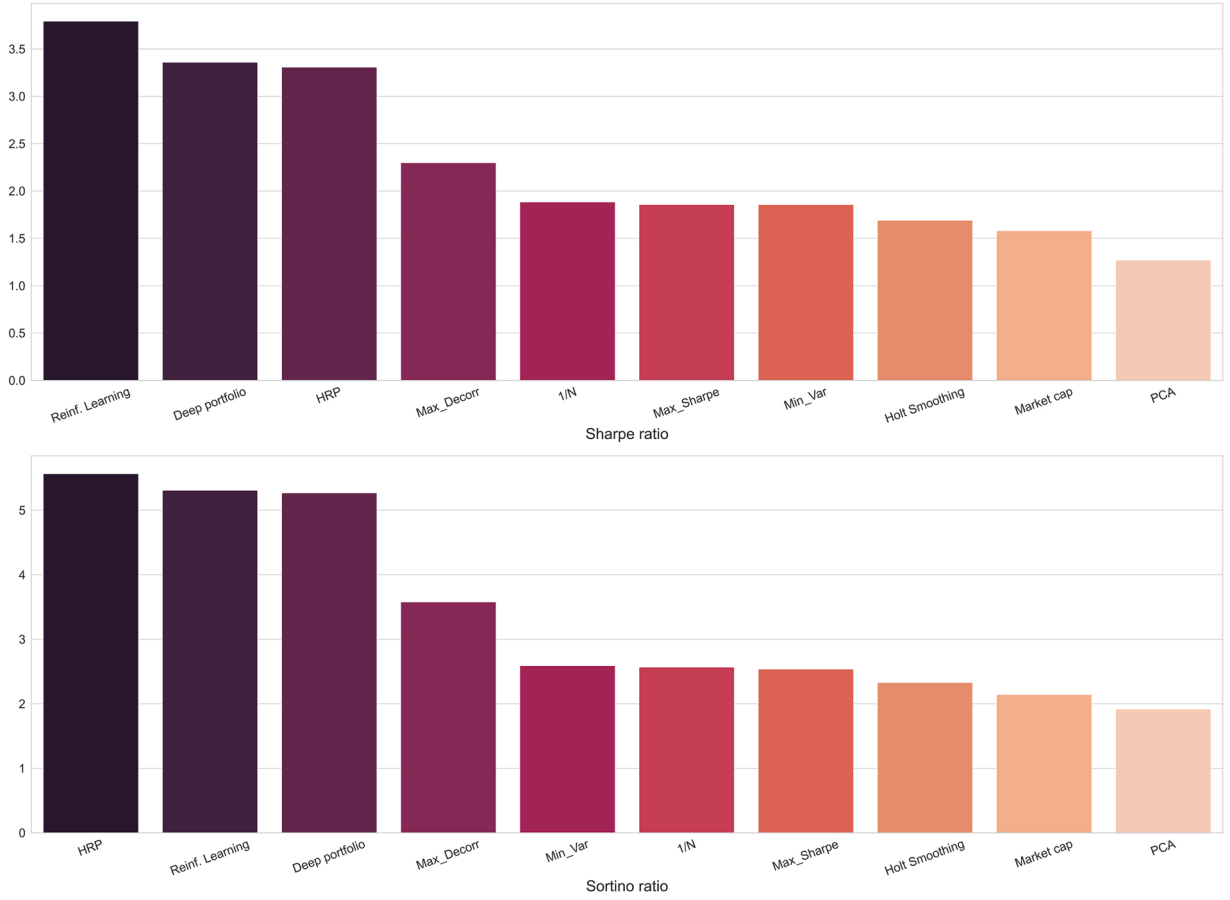

Fig. 3. Sharpe and Sortino ratios for VN30 experiment.

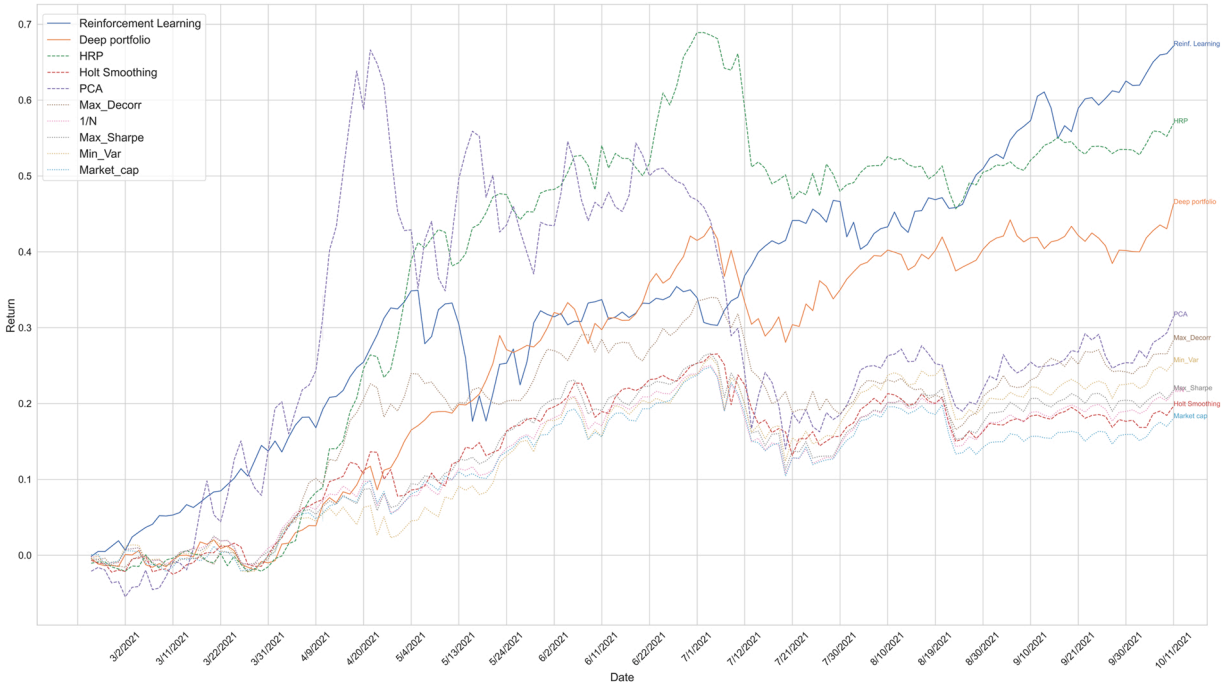

Fig. 4. Cumulative returns of VN30 experiment.

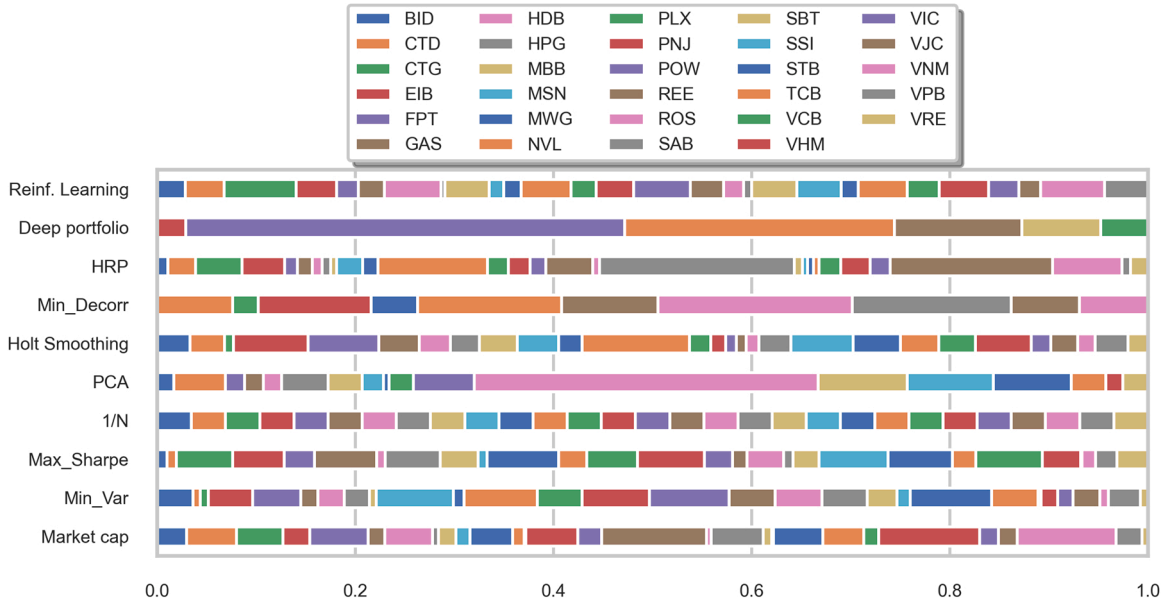

Fig. 5. Assets' weight allocation of VN30 experiment.

compared to the baseline models, but the magnitude is not significant. In particular, reinforcement learning outperforms optimization models (PCA, HRP, and Holt's smoothing) in terms of both daily risk metrics (lower risk) and daily expected returns (higher returns).

As can be seen in Table 2, the reinforcement learning model has the highest HPR of 3.7, and the deep portfolio models rank second, with 3.3. In terms of the Sortino ratio, the HRP portfolio outperformed the reinforcement learning portfolio. The Sharpe ratio of all portfolios is larger than 1; however, for baseline portfolios, PCA, and the Holt's smoothing portfolio, Sharpe ratios are about 1.7 on average, approximately half of the values from the reinforcement learning portfolio.

Additionally, Figs. 2 and 3 illustrate the results of Table 2 using plots of daily returns, the Sharpe ratio, and the Sortino ratio. It is observed from the daily return distribution in Fig. 2 that the reinforcement learning portfolio has lower volatility and higher daily

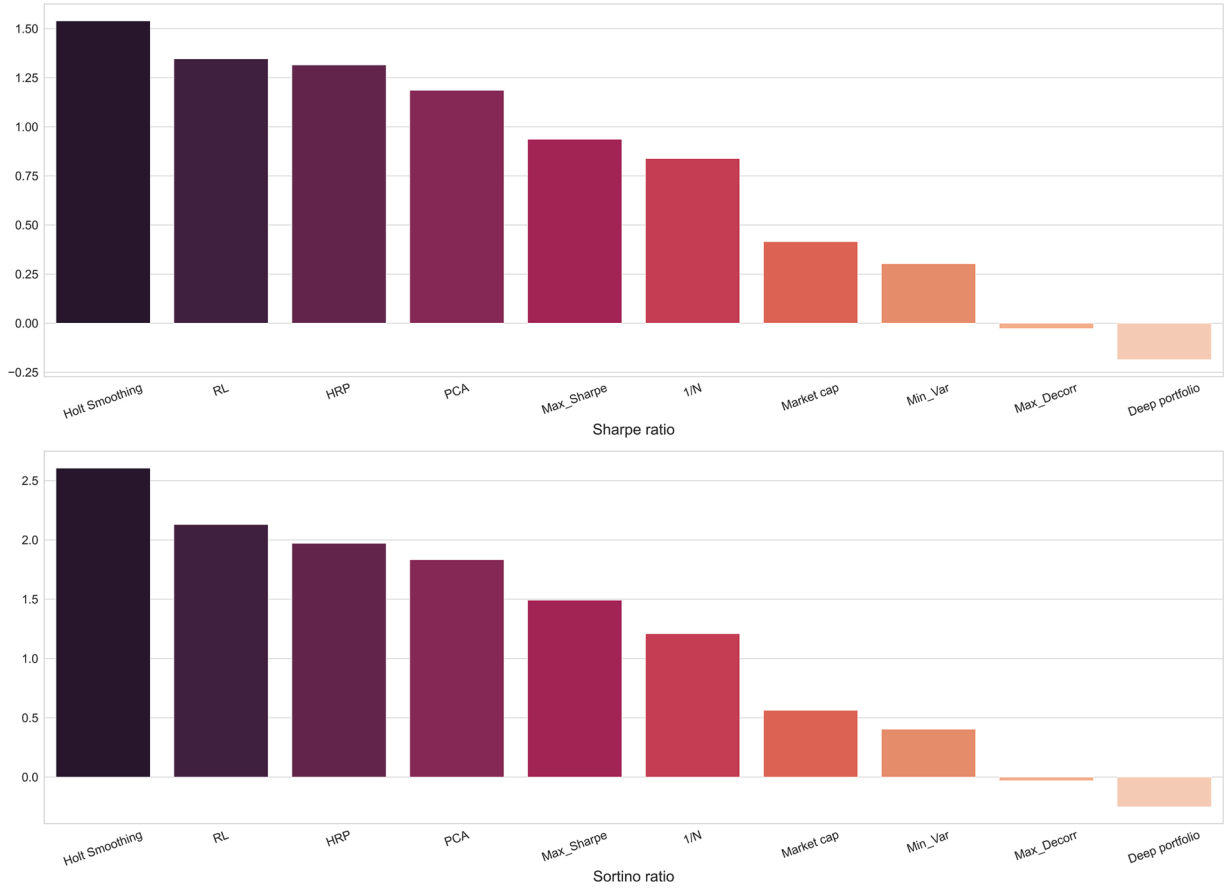

Fig. 6. Sharpe and Sortino ratios of ETF portfolio experiment.

median return than the baseline and optimization models. Deep models also have significantly higher median returns but slightly higher volatility than the baseline models.

In summary, reinforcement learning and deep portfolios outperform all mean-variance baseline models, PCA, and Holt's smoothing in terms of return characteristics, risk characteristics, and risk-reward metrics. Additionally, the HRP portfolio can overcome the drawbacks of mean-variance models and provide reliable results for portfolio construction.

To emphasize the superiority of the proposed reinforcement learning models, we present their performance results in a different form. Fig. 4 shows the cumulative returns for each model. Compared to the benchmarks, the cumulative return on the reinforcement learning portfolio is the highest. For instance, the cumulative return on the reinforcement learning portfolio is 0.67, compared to 0.56 on the HRP portfolio, 0.46 on the deep learning model, 0.31 on the PCA models, 0.19 on the Holt's smoothing model, 0.28 on the Max_Decorr model, 0.21 on the 1/N and Max_Sharpe ratio models, 0.25 on the Min_Var model, and finally 0.18 on the market cap weighted model.

To better understand the strategies applied by each model, the weight allocations for each stock in the lists of VN30 are shown in Fig. 5. The reinforcement learning portfolio does not give a dominant weight to any stock, but instead focuses on diversification goals to reduce risk. In contrast, the deep portfolio, HRP, and PCA focus more on selecting a small number of expected good stocks and assigning them substantially higher weights. The weight allocation process in the baseline models also focuses on diversification goals. However, because of their lack of adaptive ability, their ability to produce good returns is very limited compared to reinforcement learning. In summary, this experimental design shows that the reinforcement learning model has the distinctive ability to maintain a well-diversified portfolio to reduce risks while simultaneously choosing the best weight allocations to maximize portfolio returns.

4.2. .Experiment 2: ETF markets (NYSE Arca)

To test whether our models performed consistently in different contexts and timeframes, in the second experimental design, we used data from 30 ETFs from the NYSE Arca from March 2016 to October 2016.

Table 3 presents the performance evaluation metrics for different portfolios with 30 ETFs. This period of time is not optimal for trading, as mean daily returns are marginally positive, with a higher value of 0.0009 from the PCA and Holt's smoothing portfolios. The reinforcement learning portfolio comes third, with 0.0007. The mean returns of these three portfolios are substantially higher than

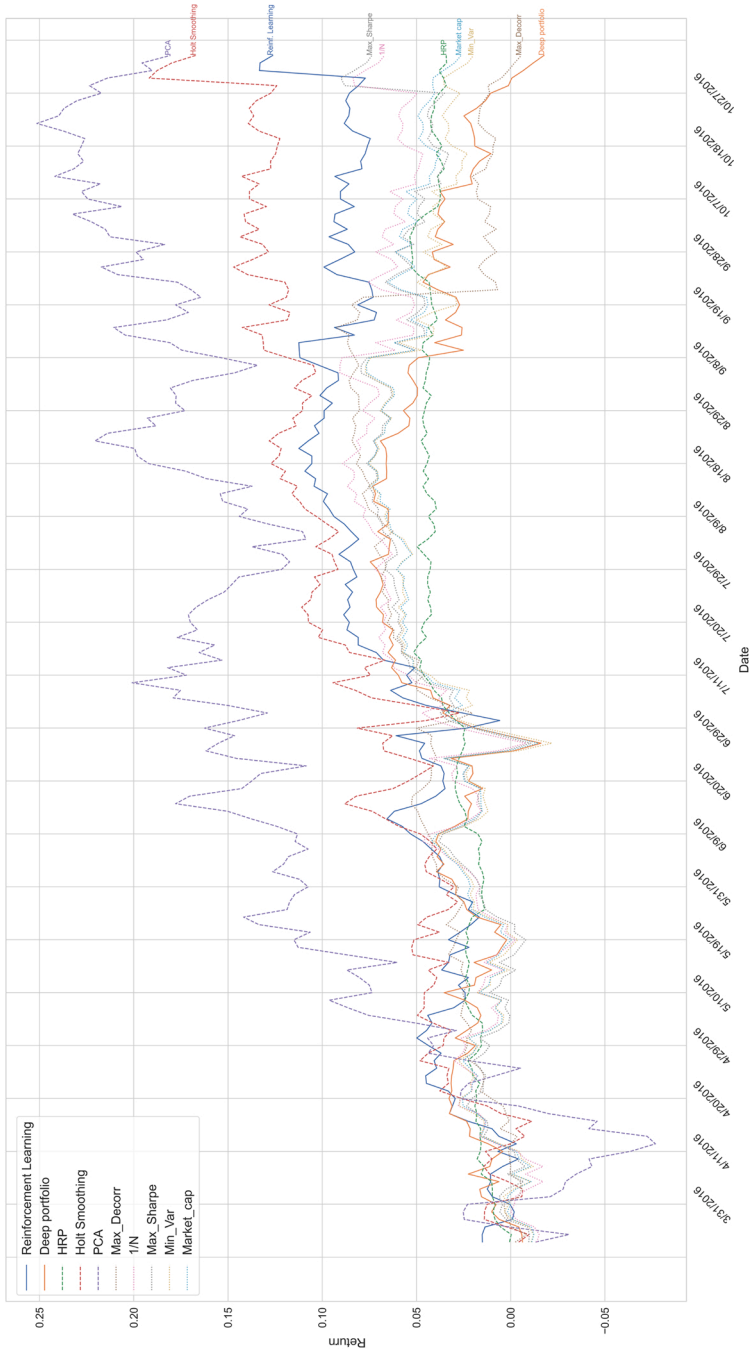

Fig. 7. Cumulative returns of ETF portfolio experiment.

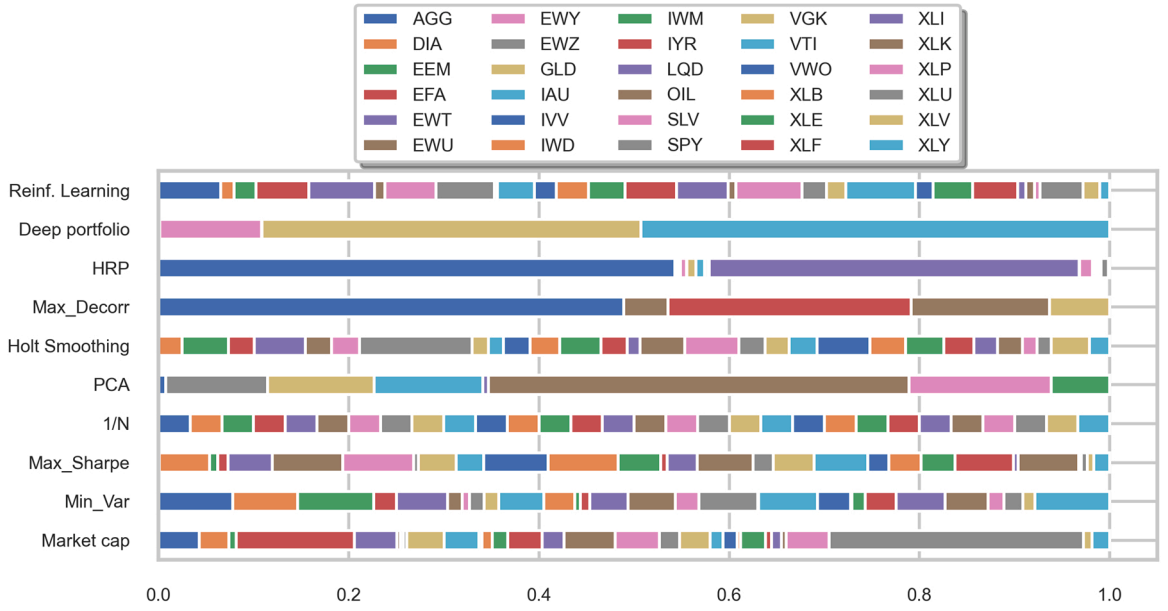

Fig. 8. Assets' weight allocation for ETF portfolio experiment.

Table 4
Deep learning portfolio evaluation metrics for VN30 portfolio.

	Reinf. learning	Deep portfolio	DeepDow portfolio	Keynes portfolio	Thorpnet portfolio
Sharpe ratio	3.7928	3.3591	1.8914	2.2002	2.5003
Sortino ratio	5.3098	5.2693	2.5742	3.0524	3.6815
E(R)	0.0032	0.0024	0.0011	0.0015	0.0017
Skewness	5.5474	0.5265	1.68	1.8577	1.1638
Kutoris	-1.4859	-0.3997	-0.926	-0.9187	-0.5434
Std(R)	0.2184	0.1821	0.151	0.176	0.1812
Max DD	-0.128	-0.1067	-0.1117	-0.1128	-0.1033
Var (1%)	-0.0193	-0.0164	-0.0145	-0.0167	-0.017
Range	0.0032	0.0024	0.0011	0.0015	0.0017
Maximum	5.5474	0.5265	1.68	1.8577	1.1638
Minimum	-1.4859	-0.3997	-0.926	-0.9187	-0.5434

Note: This table shows the evaluation metrics for the VN30 portfolios by reinforcement learning, deep portfolio models, RNN models (DeepDow and Keynes) and ANN model (Thorpnet). Regarding the evaluation metric, E(R) is mean daily return while Std(R), Max DD, Var (1%) represent for daily risk metrics. While Sortino ratio assesses the downside risk of an investment's return, Sharpe ratio calculates returns based on total market volatility, which takes both upside and downside risks into account. The other statistical measures are used to describe and summarize data. In specific, range is the difference between the largest value (maximum) and smallest value (minimum) in a set of data. While Skewness measures the asymmetry of a distribution, Kurtosis shows the flatness or peakedness of a distribution.

those of the other portfolios, with values varying from − 0.0001–0.0004. In particular, the deep portfolio performs poorly in this experiment, with the lowest mean returns. This raises the question of the stability of deep learning models in optimizing portfolio performance under unfavorable market conditions.

With regard to the daily risk characteristics, we could see that daily volatility varies from the lowest value of 0.0120 (HRP portfolio) to the highest value of 0.0292 (PCA portfolio). Again, although it has a substantially high mean return, the PCA portfolio also has exceptionally high volatility. The same pattern was observed for Max DD and var at risk (1%). Even in unfavorable market conditions, reinforcement learning portfolios maintain a solid risk-reward ratio compared to baseline models and optimization models with marginally higher volatility but substantially higher returns.

As can be seen in Table 3, the reinforcement learning model has the second-highest Sharpe ratio of 1.34 after the highest value of 1.53 from the Holt's smoothing portfolio. In addition to the two leading portfolios, HRP (1.31) and PCA (1.18) are the only two portfolios with Sharpe ratios larger than the benchmark value of 1 (Fig. 6). The Sharpe ratio of the remaining portfolios is lower than 1, with an average value of approximately 0.38, which is three times lower than the values from the reinforcement learning portfolio.

The cumulative returns of each model for the ETF markets are shown in Fig. 7. In comparison to the benchmarks, the cumulative return on the reinforcement learning portfolio is the third highest after the PCA and Holt's smoothing portfolios. Even in the period of a downtrend in the majority of ETFs from August 2016 in the sample (presented by the equal-weighted portfolio 1/N), the reinforcement

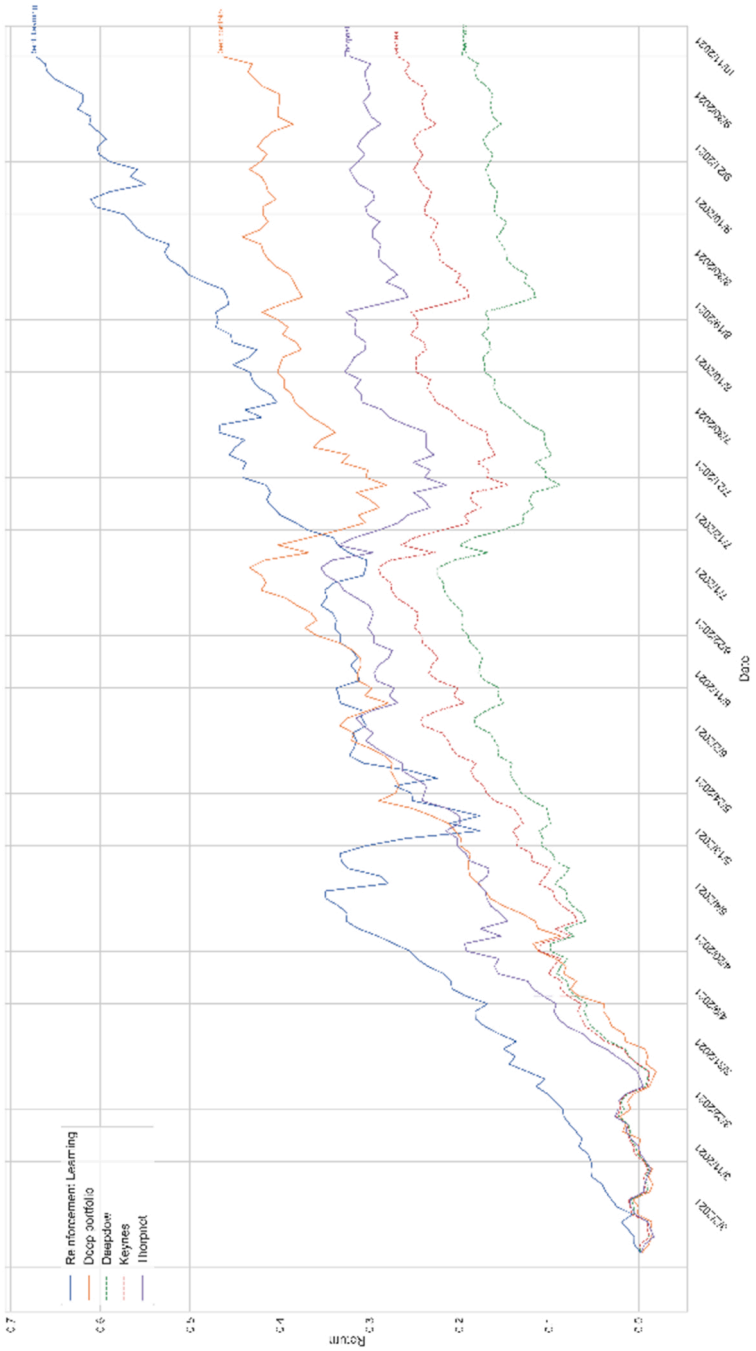

Fig. 9. Cumulative returns of VN30 portfolio experiment for different deep learning models.

Table 5
Deep learning portfolio evaluation metrics for ETF portfolio.

	Reinf. learning	Deep portfolio	DeepDow portfolio	Keynes portfolio	Thorpnet portfolio
Sharpe ratio	1.3767	-0.1900	0.5684	0.9134	-0.0076
Sortino ratio	2.1789	-0.2568	0.8131	1.3388	-0.0105
E(R)	0.0007	-0.0001	0.0001	0.0004	0.0000
Skewness	9.7627	3.2931	2.1232	6.3556	2.9379
Kutoris	0.7579	-0.6018	-0.2633	-0.0426	-0.4947
Std(R)	0.1329	0.1075	0.0447	0.119	0.0783
Max DD	-0.0562	-0.0858	-0.0391	-0.0533	-0.0616
Var (1%)	-0.013	-0.0112	-0.0045	-0.0119	-0.0081
Range	0.0007	-0.0001	0.0001	0.0004	0
Maximum	9.7627	3.2931	2.1232	6.3556	2.9379
Minimum	0.7579	-0.6018	-0.2633	-0.0426	-0.4947

Note: This table shows the evaluation metrics for the ETF portfolios by reinforcement learning, deep portfolio models, RNN models (DeepDow and Keynes) and ANN model (Thorpnet). Regarding the evaluation metric, E(R) is mean daily return while Std(R), Max DD, Var (1%) represent for daily risk metrics. While Sortino ratio assesses the downside risk of an investment's return, Sharpe ratio calculates returns based on total market volatility, which takes both upside and downside risks into account. The other statistical measures are used to describe and summarize data. In specific, range is the difference between the largest value (maximum) and smallest value (minimum) in a set of data. While Skewness measures the asymmetry of a distribution, Kurtosis shows the flatness or peakedness of a distribution.

learning portfolio shows decent performance and protects most of the returns earned from the previous period.

Fig. 8 depicts the weight allocation for each ETF in the sample. Consistent with previous experiments in the Vietnamese stock market, the reinforcement learning portfolio emphasizes diversification goals to reduce risk and does not give a dominant weight to any ETF. The same strategies were used by the best-performing portfolio in this experiment (Holt's smoothing). Baseline models also use this strategy to reduce risk, as they have the smallest volatility of return (Table 3). However, they failed to earn above-average returns. Deep learning models (deep portfolio) with a focus on selecting a number of best assets only performed well in favorable market conditions in Experiment 1 but performed poorly in unfavorable market conditions in Experiment 2.

In conclusion, the two experimental designs show that the reinforcement learning model has a consistent performance in terms of reducing risks while simultaneously maximizing portfolio returns in different asset classes and market conditions. The results suggest that a well-diversified portfolio can outperform models that focus on predicting a small number of good assets and give them substantially higher weights. However, to simultaneously reduce risk and achieve higher returns, a diversified portfolio is insufficient (baseline models). They needed to have the ability to adapt to market conditions and find the best weight allocation as the reinforcement learning model in both experiments.

4.3. Robustness tests

To assess the robustness of reinforcement learning portfolio compared to other deep learning models in varied market settings, different experiments were conducted. See Appendix 1 and 2 for the discussion of the deep learning models used in the robustness test.

4.3.1. Deep learning models

4.3.1.1. VN30 portfolio. Reinforcement learning portfolio performance is still the best compared to the different deep learning models in the VN30 portfolio regarding both portfolio cumulative returns and the Sharpe ratio (Table 4 and Fig. 9). Compared to other deep learning models, the LSTM algorithm in the deep portfolio models performs much better than other RNN models (DeepDow and Keynes) and ANN model (Thorpnet).

4.3.1.2. ETF portfolio. The reinforcement learning model is also leading in both returns and the Sharpe ratio compared to other deep learning models, even in the downtrend in the ETF experiment (Table 5 and Fig. 10). However, the LSTM algorithm in the deep portfolio is the worst compared to simple RNN models (Keynes and DeepDow).

4.3.2. Different market settings

4.3.2.1. VN30 portfolio before and after the COVID-19 pandemic. The main experience of the VN30 portfolio is from February to October 2021 when the COVID-19 pandemic was at its peak in Vietnam. To exclude the effect of the pandemic, this robustness check conducts the experiences for the VN30 portfolio before the COVID-19 outbreak (April 2019 to December 2019) as well as after Vietnam controlled the outbreak (November 2021 to August 2022).

The consistency of the reinforcement learning model compared to benchmarks methods and deep learning models is shown in Fig. 11 with different market settings. In addition, for the Vietnamese stock market, deep learning models using LSTM also perform much better than conventional RNN algorithms and benchmark models across varied market conditions (see Appendix 3 for evaluation metrics).

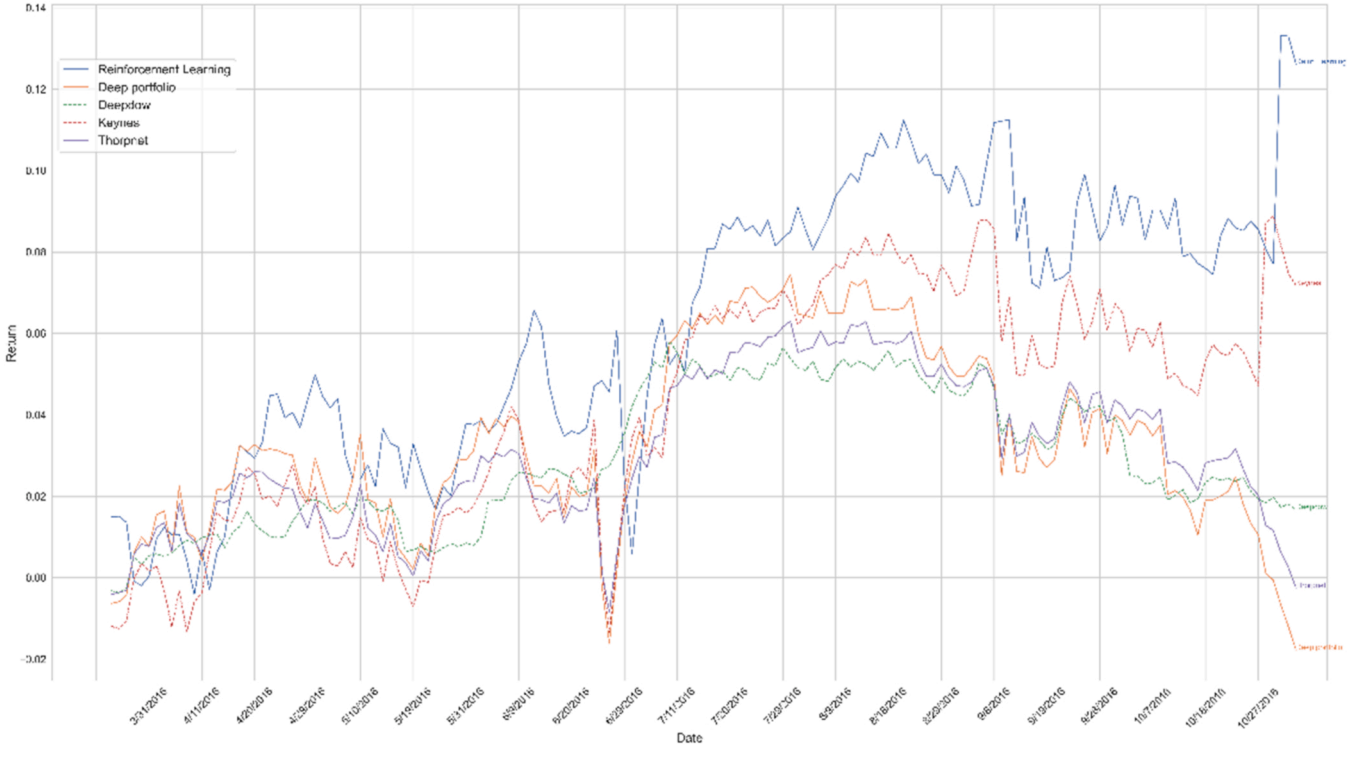

Fig. 10. Cumulative returns of ETF portfolio experiment for different deep learning models.

Fig. 11. Cumulative returns of VN30 portfolio experiment before and after the peak of the COVID-19 pandemic.

4.3.2.2. ETF portfolio in the COVID-19 pandemic. We choose the two testing samples for the COVID-19 pandemic period in the US. From February 2021 to October 2021, the growths number of new infected cases were quite stable in the US. In the period from November 2021 to August 2022, there was the peak of the COVID-19 pandemic in the US (during January and February 2022), then the pandemic was controlled later on. In both COVID-19 pandemic situations, the reinforcement learning model shows the strongest results regarding portfolio performances and the Sharpe ratio (Fig. 12 and Appendix 4). In contrast to the performances in emerging stock markets like Vietnam, deep learning models using LSTM and a conventional RNN model did not outperform benchmark models in developed stock markets even when controlling for different market settings.

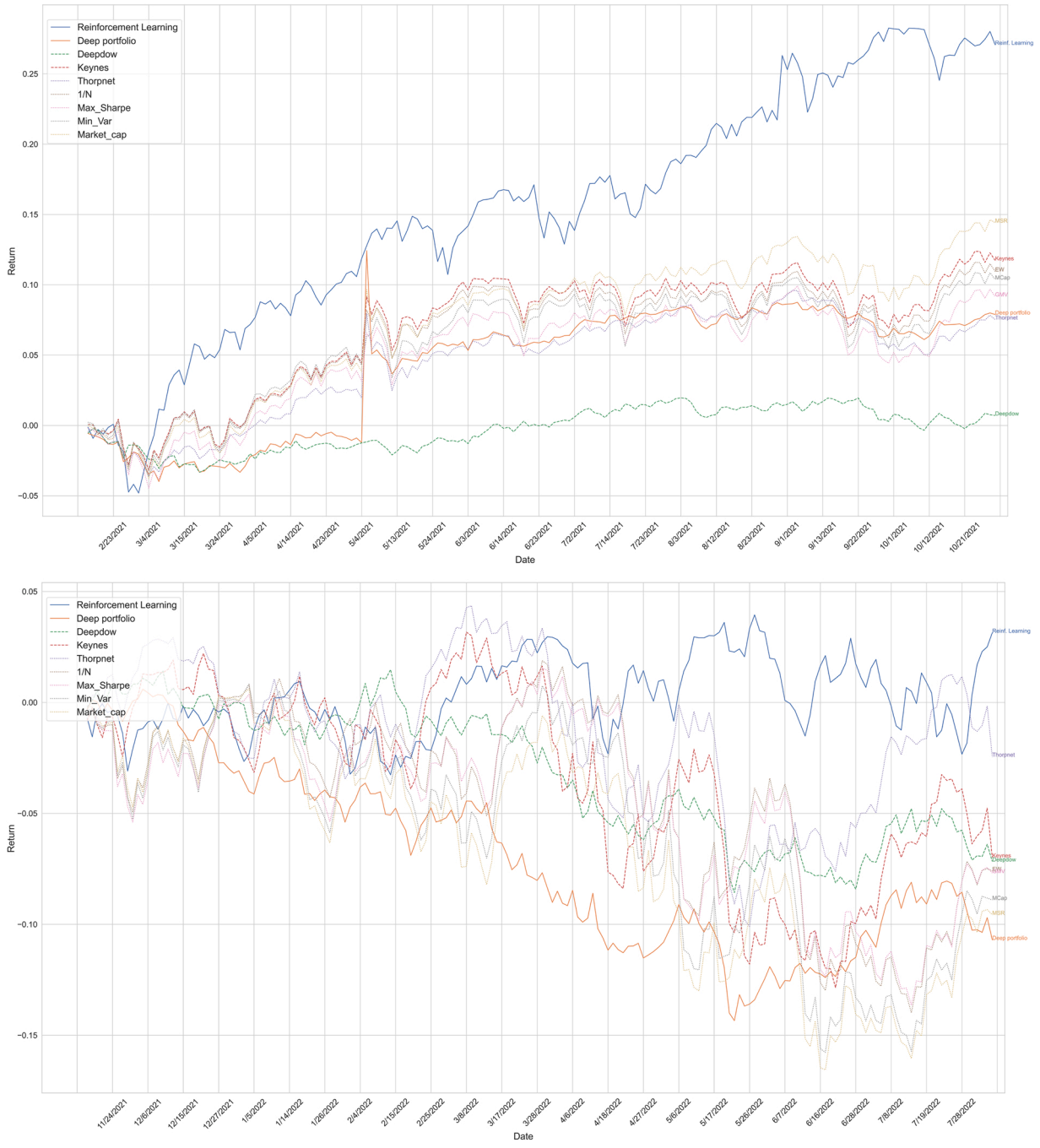

Fig. 12. Cumulative returns of ETF portfolio experiment during and after the peak of the COVID-19 pandemic.

5. Discussion and conclusion

5.1. Discussion of key findings

In this study, we used deep learning and reinforcement learning models to directly optimize the Sharpe ratio of a portfolio. Advanced machine learning models omit the typical forecasting phase and enable optimization of portfolio weights using the gradient ascent to update the model parameters. We tested whether advanced machine learning techniques could provide a better methodology than traditional methods in terms of portfolio risk management. Rather than utilizing only individual assets to construct a portfolio, we also used exchange-traded funds (ETFs) representing market indices to construct a second experiment to explore the performance of different portfolio construction methods in distinct contexts. In Experiment 1, 29 Vietnamese stocks were utilized to construct a

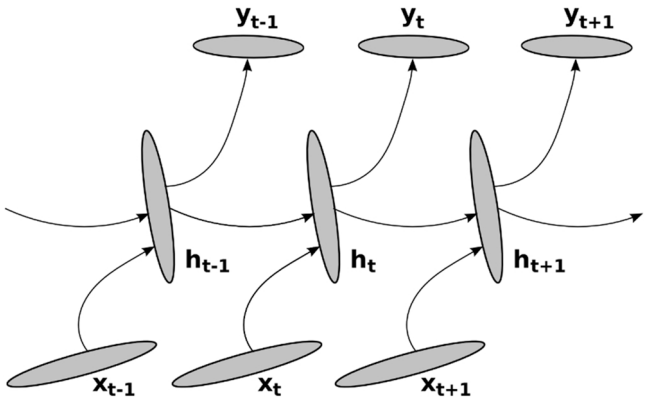

Fig. A1. Conventional recurrent neuron network.
Source: Pascanu et al. (2013).

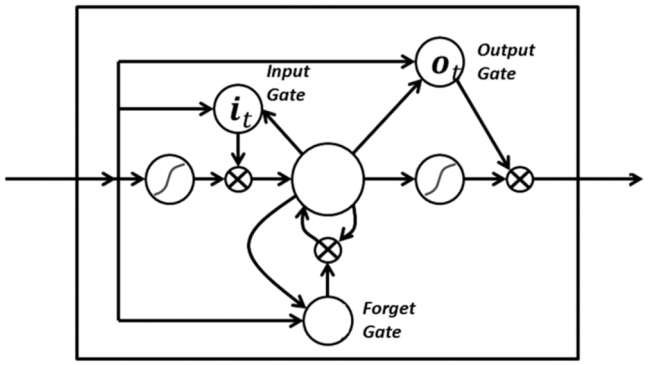

Fig. A2. LSTM memory unit.

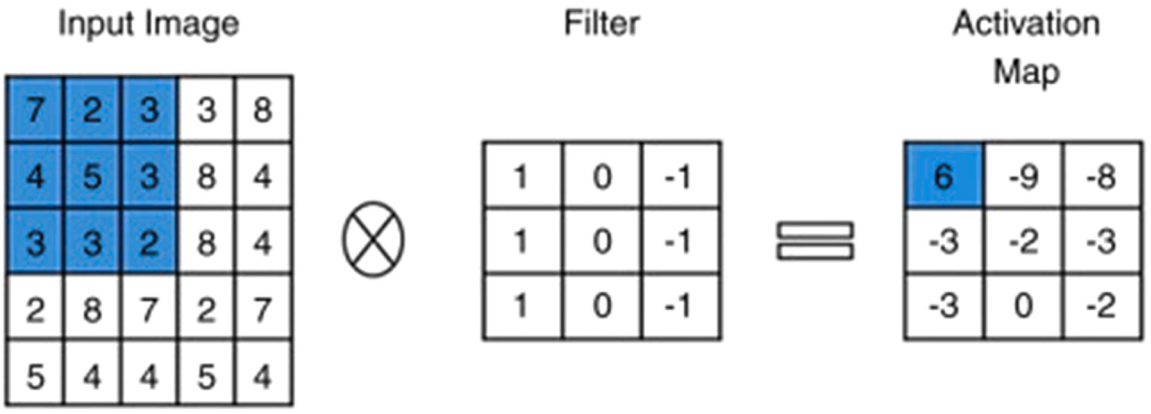

Fig. A3. Example of the convolution process.
Source: Mostafa and Wu (2021)).

portfolio in which 30 ETFs from the NYSE Arca are used.

We compared machine learning models against a variety of well-known methods, including classical mean-variance optimization (Max_Sharpe ratio), minimum variance (Min_Var), maximum decorrelation (Max_Decorr), equal-weighted (1/N), market capitalization-weighted (Market cap), and other well-known approaches such as hierarchical risk parity (HRP), principal components (PCA), and Holt's smoothing process to forecast asset price (Holt's smoothing). The two experiments in this study provided several findings.

First, in terms of cumulative returns and the Sharpe ratio, this research demonstrated that the reinforcement learning model

Table C1
Before the COVID-19 pandemic from April 2019 to December 2019.

	Reinf. Learning	Deep portfolio	Deepdow	Keynes	Thorpnet	1/N	Min_Var	Market cap	Max_Sharpe ratio
E(R)	0.001	0.000	0.000	0.000	0.001	0.000	0.000	0.000	0.000
Std(R)	0.127	0.118	0.084	0.102	0.145	0.102	0.114	0.121	0.095
Max DD	-0.066	-0.066	-0.089	-0.090	-0.088	-0.090	-0.074	-0.085	-0.084
Var (1%)	-0.012	-0.012	-0.009	-0.011	-0.014	-0.011	-0.011	-0.012	-0.010
Sharpe ratio	1.818	0.970	-1.181	-0.250	1.317	-0.284	0.707	0.533	-0.841
Sortino ratio	2.799	1.480	-1.573	-0.344	1.911	-0.390	1.042	0.755	-1.127
Range	0.001	0.000	0.000	0.000	0.001	0.000	0.000	0.000	0.000
Maximum	1.471	0.574	0.633	0.528	0.314	0.533	0.633	-0.008	0.195
Minimum	-0.026	0.128	-0.066	-0.160	-0.286	-0.162	0.009	-0.153	-0.125
Skewness	1.471	0.574	0.633	0.528	0.314	0.533	0.633	-0.008	0.195
Kurtosis	-0.026	0.128	-0.066	-0.160	-0.286	-0.162	0.009	-0.153	-0.125

Note: This table shows the evaluation metrics for the portfolios constructed from VN30 stocks before the COVID-19 pandemic from April 2019 to December 2019 by deep learning methods (deep portfolio and reinforcement learning), optimization methods (HRP, PCA, and Holt's smoothing), and baseline methods (1/N, Max_Sharpe ratio, Min_Var, Max_Decorr and Market cap weighted). Regarding the evaluation metric, E(R) is mean daily return while Std(R), Max DD, Var (1%) represent for daily risk metrics. While Sortino ratio assesses the downside risk of an investment's return, Sharpe ratio calculates returns based on total market volatility, which takes both upside and downside risks into account. The other statistical measures are used to describe and summarize data. In specific, range is the difference between the largest value (maximum) and smallest value (minimum) in a set of data. While Skewness measures the asymmetry of a distribution, Kurtosis shows the flatness or peakedness of a distribution.

Table C2
After the peak of COVID-19 pandemic from November 2021 to August 2022.

	Reinf. Learning	Deep portfolio	Deepdow	Keynes	Thorpnet	1/N	Min_Var	Market cap	Max_Sharpe ratio
E(R)	0.001	0.000	-0.001	0.000	-0.001	-0.001	-0.001	-0.001	-0.001
Std(R)	0.177	0.223	0.212	0.221	0.210	0.218	0.226	0.193	0.226
Max DD	-0.115	-0.172	-0.233	-0.223	-0.215	-0.240	-0.257	-0.201	-0.257
Var (1%)	-0.017	-0.023	-0.022	-0.023	-0.022	-0.023	-0.024	-0.021	-0.024
Sharpe ratio	1.187	0.112	-0.500	-0.249	-0.629	-0.497	-0.786	-0.787	-0.786
Sortino ratio	1.643	0.161	-0.642	-0.323	-0.806	-0.636	-0.998	-1.026	-0.998
Range	0.001	0.000	-0.001	0.000	-0.001	-0.001	-0.001	-0.001	-0.001
Maximum	2.054	1.736	2.757	2.867	3.530	2.704	1.991	2.049	1.991
Minimum	-0.691	-0.057	-0.750	-0.760	-0.794	-0.768	-0.696	-0.488	-0.696
Skewness	2.054	1.736	2.757	2.867	3.530	2.704	1.991	2.049	1.991
Kurtosis	-0.691	-0.057	-0.750	-0.760	-0.794	-0.768	-0.696	-0.488	-0.696

Note: This table shows the evaluation metrics for the portfolios constructed from VN30 stocks after the peak of COVID-19 pandemic from November 2021 to August 2022 by deep learning methods (deep portfolio and reinforcement learning), optimization methods (HRP, PCA, and Holt's smoothing), and baseline methods (1/N, Max_Sharpe ratio, Min_Var, Max_Decorr and Market cap weighted). Regarding the evaluation metric, E(R) is mean daily return while Std(R), Max DD, Var (1%) represent for daily risk metrics. While Sortino ratio assesses the downside risk of an investment's return, Sharpe ratio calculates returns based on total market volatility, which takes both upside and downside risks into account. The other statistical measures are used to describe and summarize data. In specific, range is the difference between the largest value (maximum) and smallest value (minimum) in a set of data. While Skewness measures the asymmetry of a distribution, Kurtosis shows the flatness or peakedness of a distribution.

provided the best performance in Experiment 1 and consistently good performance over traditional methods in Experiment 2 when the market condition was not as favorable. This finding is consistent with recent literature on the effectiveness of reinforcement learning in portfolio optimization. For instance, Candar and Üstündağ (2022) proved that reinforcement learning agents could bring a better weighting model to capitalization-weighted indices of stock markets, especially in emerging markets. Their study also demonstrated that their reinforcement learning agents offered superior returns and risk-adjusted returns with better Sharpe and Sortino ratios. Wu et al. (2021) conducted experiments and proved that portfolio management using reinforcement learning with the Sharpe ratio reward function surpassed the traditional return-based reward function.

In contrast, the deep learning model performed relatively well in Experiment 1 when the market condition was favorable for trading but lost its competitive edge in Experiment 2. The inconsistent performances of deep learning models were also recorded by literature such as Wang et al. (2020a), which found that even published models are sensitive to the data distribution. Recently, Ma et al. (2021) found that a simple machine learning algorithm (e.g., random forest) outperforms deep learning models (e.g., LSTM, RNN) in the tasks of portfolio optimization. This finding raises the question of whether deep learning models such as LSTM could outperform other methods given the deep learning model's sensitivity to data distribution and market conditions. Similarly, the performance of models such as HRP, PCA, and Holt's smoothing are not consistent across the two experiments with different trading conditions. They performed very well in one condition and underperformed in other conditions. This finding highlights one of the strengths of reinforcement learning models: the ability to adapt to dynamic trading conditions using dynamic programming in different market conditions and both frontier and developed financial market settings.

Second, the performance differences between the models in this study could be due to the diversification degrees used for portfolio construction. Models such as reinforcement learning, Holt's smoothing, and traditional mean-variance follow a heavy-diversification

Table D1
During the peak of COVID-19 pandemic from February 2021 to October 2021.

	Reinf. Learning	Deep portfolio	Deepdow	Keynes	Thorpnet	1/N	Min_Var	Market cap	Max_Sharpe ratio
Sharpe ratio	2.5320	0.5770	0.2081	1.3004	1.0232	1.2590	1.1315	1.5205	1.0979
Sortino ratio	3.9904	1.1593	0.2857	2.0404	1.7808	1.9199	1.6269	2.5228	1.6409
E(R)	0.0015	0.0004	0.0000	0.0006	0.0004	0.0006	0.0005	0.0008	0.0005
Std(R)	0.1497	0.1914	0.0475	0.1263	0.1031	0.1217	0.1284	0.1328	0.1170
Max DD	-0.0482	-0.0866	-0.0310	-0.0460	-0.0538	-0.0468	-0.0501	-0.0482	-0.0540
Var (1%)	-0.0140	-0.0194	-0.0049	-0.0124	-0.0103	-0.0120	-0.0127	-0.0130	-0.0116
Range	0.0015	0.0004	0.0000	0.0006	0.0004	0.0006	0.0005	0.0008	0.0005
Maximum	2.6656	98.6279	1.1419	5.7834	40.0819	2.9750	0.3516	11.9066	2.3199
Minimum	0.0919	6.7910	-0.4095	0.7726	3.6724	0.3388	-0.3234	1.5000	0.1381
Skewness	2.6656	98.6279	1.1419	5.7834	40.0819	2.9750	0.3516	11.9066	2.3199
Kurtosis	0.0919	6.7910	-0.4095	0.7726	3.6724	0.3388	-0.3234	1.5000	0.1381

Note: This table shows the evaluation metrics for the ETF portfolios constructed from the NYSE Arca during the peak of COVID-19 pandemic from February 2021 to October 2021 by deep learning methods (deep portfolio and reinforcement learning), optimization methods (HRP, PCA, and Holt's smoothing), and baseline methods (1/N, Max_Sharpe ratio, Min_Var, Max_Decorr and Market cap weighted). Regarding the evaluation metric, E(R) is mean daily return while Std(R), Max DD, Var (1%) represent for daily risk metrics. While Sortino ratio assesses the downside risk of an investment's return, Sharpe ratio calculates returns based on total market volatility, which takes both upside and downside risks into account. The other statistical measures are used to describe and summarize data. In specific, range is the difference between the largest value (maximum) and smallest value (minimum) in a set of data. While Skewness measures the asymmetry of a distribution, Kurtosis shows the flatness or peakedness of a distribution.

Table D2
After the peak of COVID-19 pandemic from November 2021 to August 2022.

	Reinf. Learning	Deep portfolio	Deepdow	Keynes	Thorpnet	1/N	Min_Var	Market cap	Max_Sharpe ratio
E(R)	0.0001	-0.0006	-0.0004	-0.0004	-0.0002	-0.0005	-0.0006	-0.0006	-0.0005
Std(R)	0.1194	0.0851	0.0820	0.1591	0.1479	0.1580	0.1724	0.1665	0.1499
Max DD	-0.0623	-0.1403	-0.0966	-0.1524	-0.1380	-0.1510	-0.1597	-0.1640	-0.1449
Var (1%)	-0.0122	-0.0094	-0.0089	-0.0169	-0.0155	-0.0168	-0.0184	-0.0178	-0.0160
Sharpe ratio	0.3683	-1.7590	-1.2230	-0.6142	-0.2313	-0.6730	-0.7229	-0.8049	-0.7183
Sortino ratio	0.5293	-2.2000	-1.5788	-0.8209	-0.3112	-0.8997	-0.9625	-1.0763	-0.9591
Range	0.0001	-0.0006	-0.0004	-0.0004	-0.0002	-0.0005	-0.0006	-0.0006	-0.0005
Maximum	0.4071	0.4116	0.6031	0.3524	0.5643	0.3620	0.1458	-0.0096	0.3850
Minimum	-0.0869	-0.4045	-0.3795	-0.3157	-0.3771	-0.3074	-0.2990	-0.2261	-0.2960
Skewness	0.4071	0.4116	0.6031	0.3524	0.5643	0.3620	0.1458	-0.0096	0.3850
Kurtosis	-0.0869	-0.4045	-0.3795	-0.3157	-0.3771	-0.3074	-0.2990	-0.2261	-0.2960

Note: This table shows the evaluation metrics for the ETF portfolios constructed from the NYSE Arca after the peak of COVID-19 pandemic from November 2021 to August 2022 by deep learning methods (deep portfolio and reinforcement learning), optimization methods (HRP, PCA, and Holt's smoothing), and baseline methods (1/N, Max_Sharpe ratio, Min_Var, Max_Decorr and Market cap weighted). Regarding the evaluation metric, E(R) is mean daily return while Std(R), Max DD, Var (1%) represent for daily risk metrics. While Sortino ratio assesses the downside risk of an investment's return, Sharpe ratio calculates returns based on total market volatility, which takes both upside and downside risks into account. The other statistical measures are used to describe and summarize data. In specific, range is the difference between the largest value (maximum) and smallest value (minimum) in a set of data. While Skewness measures the asymmetry of a distribution, Kurtosis shows the flatness or peakedness of a distribution.

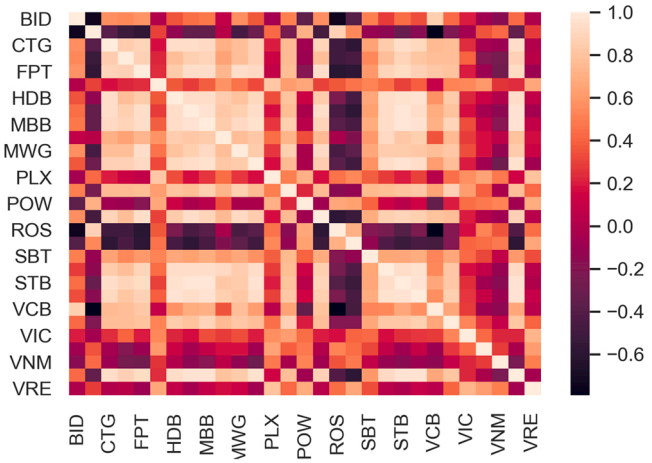

Fig. E1. Correlation heatmap between stocks in VN30 portfolio.

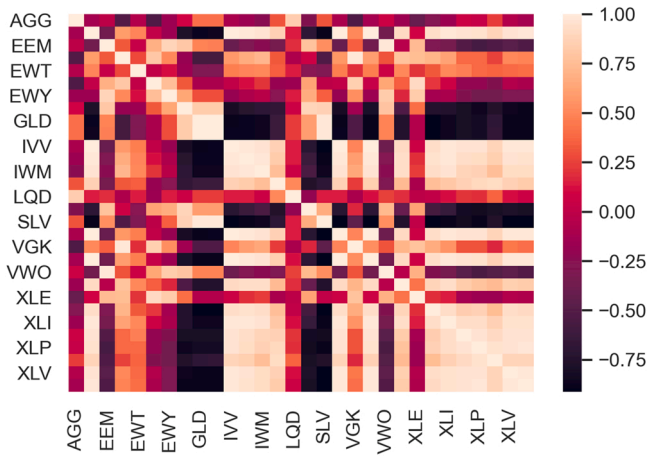

Fig. E2. Correlation heatmap between stocks in ETF portfolio.

strategy in portfolio construction. No dominant weight is assigned to a single asset or a small number of assets in the portfolios. Following this strategy, the investment portfolio consists of a relatively large number of assets. In contrast, other models, such as the deep learning model, HRP, PCA, or max decorrelation, focus on selecting a small number of assets and giving them dominant weights in portfolio constructions. Following this strategy, the performance of these portfolios depends on the performance of specific assets and cannot completely diversify unsystematic risks. Therefore, these models can perform well in some conditions and fall back in others, even for sophisticated models such as the deep learning model.

More interestingly, although following the same strategy to maximize portfolio diversification, reinforcement learning models consistently outperform traditional mean-variance models in terms of the Sharpe ratio, Sortino ratio, and cumulative returns in both experiments (Figs. 2 and 5). This highlights the fact that reinforcement learning with adaptive programming could be a game-changer in defining a new standard for the return-risk ratio and efficiency of asset allocation beyond the traditional optimization methods.

Finally, the results of the two experiments show that traditional mean-variance optimization methods of portfolio construction consistently underperform machine learning models and other statistical methods to a relatively large extent. This finding suggests that with new advancements in machine learning models, using computer algorithms in financial decisions could provide novel and consistent competitive edges in terms of both risk management and return maximization.

5.2. Limitations and future works

Although this research provides valuable insights, it has certain drawbacks. First, we want to continue this effort by examining portfolio performance using other objective functions. Given our approach's adaptable structure of reinforcement learning, we can maximize the Sortino ratio or even the degree of diversification of a portfolio, as long as functions are differentiable. Second, only data from the stock market were examined. However, owing to the diversity of economic situations, it is vital to evaluate and compare the adaptability and efficacy of portfolio construction models in various financial markets, such as bonds or cryptocurrency. Third, this study focuses on comparing the effectiveness in portfolio construction of reinforcement learning, deep learning models and other conventional methods. Thus, very little work is done on hyperparameter tuning. Therefore, future studies could focus on a few reinforcement learning and deep learning models to compare their performance with their hyperparameters being tuned. Next, this study examines distinctive models that have little in common. It would be interesting in future research to test whether the combination of these models into one framework or program could produce even better performance in terms of the risk-reward ratio. Finally, this study only compares the performance of portfolios using advanced machine learning algorithms (reinforcement learning and deep learning) with traditional approaches. Thus, future research might take this study as a starting point to further investigate whether sophisticated machine learning algorithms are more successful in developing markets or developed markets, considering their huge differences in market settings, in order to provide more meaningful findings.

CRediT authorship contribution statement

Vu Minh Ngo: Conceptualization, Data curation, Investigation, Writing – original draft, Validation, Project administration, Funding acquisition, Supervision. **Huan Huu Nguyen:** Formal analysis, Methodology, Software, Writing – original draft, Visualization, Investigation, Writing – review & editing. **Phuc Van Nguyen:** Writing – original draft, Resources, Validation.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to

influence the work reported in this paper.

Data Availability

Data will be made available on request.

Acknowledgement

This research is funded by University of Economics Ho Chi Minh City, Vietnam (UEH).

APPENDICES.

Appendix A. Recurrent neural network

Appendix Fig. A1.

Different models of deep learning are used in this study, including the deep portfolio model, the Deepdow model, the Thorpnet model and the Keynes model. Because of the nature of the sequential input of financial data in this study, Recurrent Neural Network (RNN) is usually referred to as one of the state of the art algorithm for many applications, including financial prediction problems (Abiodun et al., 2018). Due to its internal memory, it is the first algorithm to recall its input, making it ideal for machine learning issues involving sequential data. Compared to other algorithms, recurrent neural networks may develop a far deeper grasp of a sequence and its environment. Thus, the deep learning models in this study are based on the basics of the conventional RNN algorithm.

A recurrent neural network (RNN) replicates a discrete-time dynamical system with an input x_t, an output y_t, and a hidden state h_t. In our notation, time is represented by the subscript t. The dynamic system is described as:

$$\left\{ \begin{array}{c} h_t = f_h(x_t, h_{t-1}) \\ y_t = f_o(h_t) \end{array} \right\}$$

where state transition function and an output function are represented as f_h and f_o, accordingly. The set of parameters, θ_h and θ_o for functions f_h and f_o could be estimated by minimizing the following cost function:

$$J(\theta) = \frac{1}{N} \sum_{n=1}^{N} \sum_{t=1}^{Tn} d(y_t^{(n)}, f_o(h_t^{(n)})) \pm$$

where $h_t^{(n)} = f_h(x_t^{(n)}, h_{t-1}^{(n)})$ and $h_0^{(n)} = 0$. $D(y_t^{(n)}, f_o(h_t^{(n)}))$ is the errors between the estimated value and observed value of $y_t^{(n)}$..

Appendix B. Deep learning models

Deep portfolio model

The Deep portfolio model used in the experiments uses the long short-term memory (LSTM) algorithm and architectures for their learning. Long short-term memory networks (LSTMs) are an extension of Recurrent neural networks (RNNs) that, in essence, expand the memory. LSTMs allow RNNs to retain inputs for an extended length of time. This is because of LSTMs store information in a memory, similar to computer memory. The LSTM is capable of reading, writing, and erasing data from its memory using its LSTM unit (Fig. A2) in each time step. A number of different layers of memory units are stacked together, which are called hidden layers of neural networks in LSTM models (Fig. 1).

: Forget gate. The sigmoid function (output is $\in [0, 1]$) decides whether the part of the old output is necessary to feed into the calculation for the value of the current state.

: Input gate. The sigmoid function (output is $\in [0, 1]$) decides which information from the current state is important and which one is not important.

: Output gate. The sigmoid function (output is $\in [0, 1]$) determines the value of the next hidden state based on the inputs from current and old information.

: dot product calculation of inputs.

Deepdow model

The Deepdow model is a conventional RNN with 32 hidden layers. The Deepdow model uses a one-dimensional convolutional layer to transform input data into feature maps for RNN processing. Deepdow model using the convex optimizers to find the weight

allocation for each stock in the portfolio.

A convolutional layer is the fundamental component of a CNN. It includes a collection of filters (or kernels) whose parameters must be learned during training. Typically, the size of the filter is less than the size of the input data itself. Each filter convolutionally processes the input data to generate an activation map. For convolution, the filter is slid over the input data dimension, and at each spatial point, the dot product between each element of the filter and the input is computed. Fig. A3 is an illustration of the convolution process. The first item of the activation map (highlighted in blue in Fig. A3) is computed by convolving the filter with the input section highlighted in blue. This procedure is repeated for each element in the input to produce the activation map. The convolutional layer's output volume is formed by stacking the activation maps of each filter along the depth dimension. Every element of the activation map may be interpreted as the neuron's output. Consequently, each neuron is linked to a tiny local region in the input picture, and the size of the region corresponds to the size of the filter. All of the neurons in an activation map have the same parameters. Due to the local connection of the convolutional layer, the network is compelled to train filters with the highest response to a local input area. Initial convolutional layers extract low-level characteristics, and subsequent layers extract high-level features.

Keynes model

The Keynes model is also a conventional RNN with 32 hidden layers. However, instead of using the convex optimizers, it uses the softmax allocators to find the weight allocation for each stock in the portfolio. The one-dimensional convolutional layer is also used to transform input data into feature maps for RNN processing.

Thorpnet model

The Thorpnet model is a typical neuron network that disregards the input as a dynamically evolving temporal tensor and RNN is not used. Training teaches all of the critical elements for portfolio allocation. This indicates that the network discovers a single optimum set of parameters for the whole training set. The most important input is the covariance matrix between assets in the portfolio.

Appendix C. VN30 portfolito evaluation metrics for different market settings

Appendix (Tables C1 and C2).

Appendix D. EFT portfolito evaluation metrics for different market settings

Appendix (Tables D1 and D2).

Appendix E. Correlation between assets

Appendix Figs. E1 and E2.

References

Abiodun, O.I., Jantan, A., Omolara, A.E., Dada, K.V., Mohamed, N.A., Arshad, H., 2018. State-of-the-art in artificial neural network applications: a survey. Heliyon 4 (11), e00938.

Anh, D.L.T., Christopher, G., 2020. The impact of the COVID-19 lockdown on stock market performance: evidence from Vietnam. J. Econ. Stud. 48 (4), 836–851.

Arnott, R., Harvey, C.R., Markowitz, H., 2019. A backtesting protocol in the era of machine learning. J. Financ. Data Sci. 1 (1), 64–74.

Arroyo, J., Corea, F., Jimenez-Diaz, G., Recio-Garcia, J.A., 2019. Assessment of machine learning performance for decision support in venture capital investments. IEEE Access 7, 124233–124243.

Bartram, Söhnke, M., Jürgen, Branke, Giuliano De, Rossi, Mehrshad, Motahari, 2021. Machine learning for active portfolio management. J. Financ. Data Sci. 3 (3), 9–30. ⟨https://jfds.pm-research.com/content/3/3/9⟩.

Bisoi, R., Dash, P.K., Parida, A.K., 2019. Hybrid variational mode decomposition and evolutionary robust kernel extreme learning machine for stock price and movement prediction on daily basis. Appl. Soft Comput. 74, 652–678.

Bollerslev, T., 1986. Generalized autoregressive conditional heteroskedasticity. J. Econ. 31 (3), 307–327.

Bolognesi, E., Torluccio, G., Zuccheri, A., 2013. A comparison between capitalization-weighted and equally weighted indexes in the European equity market. J. Asset Manag. 14 (1), 14–26.

Campbell, J.Y., Lettau, M., Malkiel, B.G., Xu, Y., 2001. Have individual stocks become more volatile? An empirical exploration of idiosyncratic risk. J. Financ. 56 (1), 1–43.

Candar, Mert, Üstündağ, Alp, 2022. Equity portfolio optimization using reinforcement learning: emerging market case. Intell. Fuzzy Syst. 131–139. ⟨https://link.springer.com/chapter/10.1007/978-3-031-09176-6_16⟩.

Chen, R., Liang, C.Y., Hong, W.C., Gu, D.X., 2015. Forecasting holiday daily tourist flow based on seasonal support vector regression with adaptive genetic algorithm. Appl. Soft Comput. 26, 435–443.

Chong, E., Han, C., Park, F.C., 2017. Deep learning networks for stock market analysis and prediction: Methodology, data representations, and case studies. Expert Syst. Appl. 83, 187–205.

Christoffersen, P., Jacobs, K., Mimouni, K., 2010. Volatility dynamics for the S&P500: Evidence from realized volatility, daily returns, and option prices. Rev. Financ. Stud. 23 (8), 3141–3189.

De Prado, M.L., 2016. Building diversified portfolios that outperform out of sample. J. Portf. Manag. 42 (4), 59–69.

DeMiguel, V., Garlappi, L., Uppal, R., 2009. Optimal versus naive diversification: how inefficient is the 1/N portfolio strategy? Rev. Financ. Stud. 22 (5), 1915–1953.

Elton, E.J., Gruber, M.J., 1977. Risk reduction and portfolio size: an analytical solution. J. Bus. 50 (4), 415–437.

Elton, E.J., Gruber, M.J., 1997. Modern portfolio theory, 1950 to date. J. Bank. Financ. 21 (11–12), 1743–1759.

Engle, R.F., 1982. Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation. Écon.: J. Econ. Soc. 987–1007.

Giang, N.K. 2020. From Extreme Turmoil, Vietnam Stocks Become World's Best. Bloomberg. https://www.bloomberg.com/news/articles/2020–04-14/from-extreme-turmoil-vietnam-stocks-become-world-s-best (August 28, 2022).
Giang, N.K., and L. Yap. 2020. Inside the Best Asian Stock Rally of May. Bloomberg. https://www.bloomberg.com/news/articles/2020–05-27/inside-asia-s-best-stock-rally-in-may-vietnam-markets-primer (August 28, 2022).
Gu, S., Kelly, B., Xiu, D., 2020. Empirical asset pricing via machine learning. Rev. Financ. Stud. 33 (5), 2223–2273.
Holt, Charles C., 1957. Forecasting trends and seasonal by exponentially weighted averages. Int. J. Forecast. 20 (1), 5–10. https://doi.org/10.1016/j.ijforecast.2003.09.015.
Jurczenko, E. (Ed.), 2020. Machine Learning for Asset Management: New Developments and Financial Applications. John Wiley & Sons.
Kaczmarek, T., Perez, K., 2021. Building portfolios based on machine learning predictions. Econ. Res. -Èkon. Istraživanja 1–19.
Khushi, Matloob, and Terry Lingze Meng 2019. Reinforcement Learning in Financial Markets. Data 2019, Vol. 4, Page 110 4(3): 110. https://www.mdpi.com/2306–5729/4/3/110/htm (August 29, 2022).
Kolm, P.N., Tütüncü, R., Fabozzi, F.J., 2014. 60 years of portfolio optimization: practical challenges and current trends. Eur. J. Oper. Res. 234 (2), 356–371.
Kolm, Petter N., Ritter, Gordon, 2019. Dynamic replication and hedging: a reinforcement learning approach. J. Financ. Data Sci. 1 (1), 159–171. ⟨https://jfds.pm-research.com/content/1/1/159⟩.
Kritzman, M., Page, S., Turkington, D., 2010. In defense of optimization: the fallacy of 1/N. Financ. Anal. J. 66 (2), 31–39.
Lahmiri, S., Bekiros, S., 2020. Intelligent forecasting with machine learning trading systems in chaotic intraday Bitcoin market. Chaos Solitons Fractals 133, 109641.
Li, Yuxi. 2017. Deep Reinforcement Learning: An Overview." https://arxiv.org/abs/1701.07274v6 (August 29, 2022).
Li, J., Monroe, W., Ritter, A., Galley, M., Gao, J., & Jurafsky, D. (2016). Deep reinforcement learning for dialogue generation. arXiv preprint arXiv:1606.01541.
Li, Yaoming, Junfeng Wu, and Yun Chen. 2020. Asset Allocation Based on Reinforcement Learning. IEEE International Conference on Industrial Informatics (INDIN) 2020-July: 397–402.
Liang, Z., Chen, H., Zhu, J., Jiang, K., & Li, Y. (2018). Adversarial deep reinforcement learning in portfolio management. *arXiv preprint arXiv:1808.09940*.
Long, W., Lu, Z., Cui, L., 2019. Deep learning-based feature engineering for stock price movement prediction. Knowl. -Based Syst. 164, 163–173.
Ma, Y., Han, R., Wang, W., 2021. Portfolio optimization with return prediction using deep learning and machine learning. Expert Syst. Appl. 165, 113973.
Malladi, R., Fabozzi, F.J., 2017. Equal-weighted strategy: Why it outperforms value-weighted strategies? Theory and evidence. J. Asset Manag. 18 (3), 188–208.
Markowitz, H.M., 1952. Portfolio selection. J. Financ. 15–30. https://doi.org/10.2307/2975974.
Markowitz, H.M., 1991. Foundations of portfolio theory. J. Financ. 46 (2), 469–477.
McNeil, A.J., Frey, R., Embrechts, P., 2015. Quantitative Risk Management: Concepts, Techniques and Tools-revised Edition. Princeton University Press.
Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A.A., Veness, J., Bellemare, M.G., Hassabis, D., 2015. Human-level control through deep reinforcement learning. Nature 518 (7540), 529–533.
Mokkelbost, P.B., 1971. Unsystematic risk over time. J. Financ. Quant. Anal. 6 (2), 785–796.
Moody, J., Wu, L., Liao, Y., Saffell, M., 1998. Performance functions and reinforcement learning for trading systems and portfolios. J. Forecast. 17 (5–6), 441–470.
Mostafa, S., Wu, F.X., 2021. Diagnosis of autism spectrum disorder with convolutional autoencoder and structural MRI images. Neural Engineering Techniques for Autism Spectrum Disorder. Academic Press, pp. 23–38.
Partovi, M.H., Caputo, M., 2004. Principal portfolios: recasting the efficient frontier. Econ. Bull. 7 (3), 1–10.
Pascanu, R., Gulcehre, C., Cho, K., & Bengio, Y. (2013). How to construct deep recurrent neural networks. arXiv preprint arXiv:1312.6026.
Saltiel, D., Benhamou, E., Ohana, J. J., Laraki, R., & Atif, J. (2020). Drlps: Deep reinforcement learning for portfolio selection. *ECML PKDD Demo track*.
Sepp, H., Jürgen, S., 1997. Long short-term memory. Neural Comput. 9 (8), 1735–1780.
Sharpe, W.F., 1970. Portfolio Theory and Capital Markets. McGraw-Hill College.
Sharpe, W.F., 1994. The sharpe ratio. J. Portf. Manag. 21 (1), 49–58.
Silver, D., Huang, A., Maddison, C.J., Guez, A., Sifre, L., Van Den Driessche, G., Hassabis, D., 2016. Mastering the game of Go with deep neural networks and tree search. Nature 529 (7587), 484–489.
Sirignano, J., Cont, R., 2019. Universal features of price formation in financial markets: perspectives from deep learning. Quant. Financ. 19 (9), 1449–1459.
Soleymani, F., Paquet, E., 2020. Financial portfolio optimization with online deep reinforcement learning and restricted stacked autoencoder—DeepBreath. Expert Syst. Appl. 156, 113456.
Statista. 2022. Largest Stock Exchange Operators by Market Cap 2022." Statista Research Department. https://www.statista.com/statistics/270126/largest-stock-exchange-operators-by-market-capitalization-of-listed-companies/ (August 28, 2022).
Statman, M., 1987. How many stocks make a diversified portfolio? J. Financ. Quant. Anal. 22 (3), 353–363.
Sutton, R.S., & Barto, A.G. (1998). Introduction to reinforcement learning (Vol. 135). Cambridge: MIT press.
Taljaard, B.H., Maré, E., 2021. Why has the equal weight portfolio underperformed and what can we do about it? Quant. Financ. 1–14.
Tang, G.Y., 2004. How efficient is naive portfolio diversification? An educational note. Omega 32 (2), 155–160.
The Ministry of Finance. 2020. Vietnam Stock Market Has Recover in Term of Indicators, Scales and Internal Force. The Ministry of Finance. https://mof.gov.vn/webcenter/portal/vclvcstcen/pages_r/l/detailnews?dDocName=MOFUCM184378 (August 28, 2022).
Vo, Xuan, Vinh, Truong, Quang Binh, 2018. Does momentum work? Evidence from Vietnam stock market. J. Behav. Exp. Financ. 17, 10–15.
Wang, H., Zhou, X.Y., 2020. Continuous-time mean–variance portfolio selection: A reinforcement learning framework. Mathematical Finance 30 (4), 1273–1308.
Wang, Peijin, Zhang, Hongwei, Yang, Cai, Guo, Yaoqi, 2021. Time and frequency dynamics of connectedness and hedging performance in global stock markets: bitcoin versus conventional hedges. Res. Int. Bus. Financ. 58, 101479.
Wang, W., Li, W., Zhang, N., Liu, K., 2020b. Portfolio formation with preselection using deep learning from long-term financial data. Expert Syst. Appl. 143, 113042.
Wang, X., Liang, G., Zhang, Y., Blanton, H., Bessinger, Z., Jacobs, N., 2020a. Inconsistent performance of deep learning models on mammogram classification. J. Am. Coll. Radiol. 17 (6), 796–803.
Wen, Wen, Yuyu Yuan, and Jincui Yang. 2021. Reinforcement Learning for Options Trading." Applied Sciences 2021, Vol. 11, Page 11208 11(23): 11208. https://www.mdpi.com/2076–3417/11/23/11208/htm (August 29, 2022).
Wu, Mu, En, Jia Hao, Syu, Jerry Chun Wei, Lin, Ho, Jan Ming, 2021. Portfolio management system in equity market neutral using reinforcement learning. Appl. Intell. 51 (11), 8119–8131. ⟨https://link.springer.com/article/10.1007/s10489-021-02262-0⟩.
Wu, W., Chen, J., Yang, Z., Tindall, M.L., 2021. A cross-sectional machine learning approach for hedge fund return prediction and selection. Manag. Sci. 67 (7), 4577–4601.
Yu, P., Lee, J.S., Kulyatin, I., Shi, Z., Dasgupta, S., 2019. Model-based deep reinforcement learning for financial portfolio optimization. RWSDM Workshop, ICML (Vol. 1,, 2019.
Zhang, Z., Zohren, S., Roberts, S., 2019. Deeplob: Deep convolutional neural networks for limit order books. IEEE Trans. Signal Process. 67 (11), 3001–3012.
Zhang, Z., Zohren, S., Roberts, S., 2020a. Deep reinforcement learning for trading. J. Financ. Data Sci. 2 (2), 25–40.
Zhang, Z., Zohren, S., Roberts, S., 2020b. Deep learning for portfolio optimization. J. Financ. Data Sci. 2 (4), 8–20.
Zhang, Zihao, Zohren, Stefan, Roberts, Stephen, 2020. Deep reinforcement learning for trading. J. Financ. Data Sci. 2 (2), 25–40. ⟨https://jfds.pm-research.com/content/2/2/25⟩.
Zhou, F., Zhang, Q., Sornette, D., Jiang, L., 2019. Cascading logistic regression onto gradient boosted decision trees for forecasting and trading stock indices. Appl. Soft Comput. 84, 105747.