RESEARCH ARTICLE

Risk-Adjusted Deep Reinforcement Learning for Portfolio Optimization: A Multi-reward Approach

Himanshu Choudhary[1] · Arishi Orra[1] · Kartik Sahoo[1,2] · Manoj Thakur[1]

Received: 6 February 2025 / Revised: 3 May 2025 / Accepted: 14 May 2025

Abstract
Portfolio optimization is a widely studied topic in quantitative finance. Recent advances in portfolio optimization have shown promising capabilities of deep reinforcement learning algorithms to dynamically allocate funds across various potential assets to meet the objectives of prospective investors. The reward function plays a crucial role in providing feedback to the agent and shaping its behavior to attain the desired goals. However, choosing an optimal reward function poses a significant challenge for risk-averse investors aiming to maximize returns while minimizing risk or pursuing multiple investment objectives. In this study, we attempt to develop a risk-adjusted deep reinforcement learning (RA-DRL) approach leveraging three DRL agents trained using distinct reward functions, namely, log returns, differential Sharpe ratio, and maximum drawdown to develop a unified policy that incorporates the essence of these individual agents. The actions generated by these agents are then fused by employing a convolutional neural network to provide a single risk-adjusted action. Instead of relying solely on a singular reward function, our approach integrates three different functions aiming at diverse objectives. The proposed approach is tested on daily data of four real-world stock market instances: Sensex, Dow, TWSE, and IBEX. The experimental results demonstrate the superiority of our proposed approach based on several risk and return performance metrics when compared with base DRL agents and benchmark methods.

Keywords Deep reinforcement learning · Portfolio optimization · Convolutional neural network · Proximal policy optimization · Quantitative finance

1 Introduction

The primary goal of an investment is to determine an optimum allocation of funds into a pool of assets to obtain the maximum possible returns at the cost of minimum potential risk. In a real-world scenario, an investor intends to invest the available capital into various securities (mutual funds, bonds, stocks, and derivatives) to protect oneself from market volatility. The uncertain patterns present in the stock market emphasize the necessity for a well-informed investment strategy. Due to this uncertainty, determining the optimal proportion of securities becomes challenging. Such a problem is often studied in financial mathematics literature and is generally recognized as a portfolio optimization problem. The portfolio optimization problem involves allocating the available funds among the various securities to achieve the investor's objectives. Two primary goals common to most investors are risk minimization and maximization of return. Markowitz [1] was the one who first defined the trade-off between risk and returns in his seminal work. Markowitz proposed the mean-variance optimization (MVO) model, using the portfolio's mean as return and variance as a measure of risk. This idea later opened the way for research into portfolio optimization and appeared to be a milestone in modern portfolio theory [2–8]. Since the historical and

Published online: 26 May 2025

future stock returns do not always correlate, an efficient portfolio constructed using historical returns will not be efficient at all in the future.

Advancements in machine learning (ML) and deep learning (DL) algorithms persuaded researchers and practitioners to use modern learning systems for their capacity to analyze historical market data patterns and produce systematic decisions in portfolio optimization [9–14]. Perrin and Roncalli [15] demonstrated the potential benefits of using ML techniques for large-scale optimization problems and provided an alternative approach to MVO for portfolio allocation. Afterward, Chen et al. [12] introduced a hybrid approach by combining the mean-variance optimization model for portfolio selection and XGBoost for stock prediction. Pinelis and Ruppert [16] applied two random forest algorithms to build an optimal portfolio where one model predicted the excess returns while the other estimated the volatility. Zhang et al. [13] suggested a pipeline using deep-learning models that directly optimize the portfolio weights by updating the model parameters. These algorithms necessitate labeled data associated with particular objectives, such as regression or classification. However, designing the appropriate labels for financial data is challenging, particularly within portfolio management.

Reinforcement learning (RL) is a paradigm that trains an agent to act optimally in a dynamic environment to maximize the rewards. The agent learns to take optimal steps via trial and error to accumulate maximum reward [17]. Deep reinforcement learning (DRL) incorporates DL to solve the policy while dealing with complex, high-dimensional RL problems. For portfolio optimization problems, DRL provides a framework to allocate assets to maximize the investment return dynamically [18–28]. Jeong and Kim [19] presented an automated trading methodology that utilizes the deep Q-learning (DQN) [29] algorithm to identify the optimal number of stocks to trade. They also addressed the overfitting problem of DL models by a transfer learning approach. However, DQN can't be employed directly for asset allocation problems due to its limitations to discrete action space. Various researchers [18, 30–32] adopted a Proximal Policy Optimization (PPO) [33] algorithm to address the portfolio optimization problem. Jang and Seong [34] employed the Deep Deterministic Policy Gradient (DDPG) [35] method to optimize the portfolio dynamically. Liang et al. [21] utilized three continuous DRL algorithms, namely Policy Gradient (PG), PPO, and DDPG for managing portfolios and proposed a novel adversarial training approach for enhancing their training efficiency.

In RL, the reward function shapes the agent's learning process and guides its behavior toward desired outcomes. When optimizing portfolio weights using DRL techniques, the reward function is adjusted to increase portfolio return, decrease portfolio risk, or attain alternative objectives. Most of the current approaches in the literature incorporate portfolio returns or log returns as a reward to accomplish higher profits [20, 34, 36–38]. Nonetheless, this reward overlooks the associated risk, making it less practical. Various studies [25, 39, 40] addressed the portfolio optimization problem through RL by utilizing a risk-sensitive reward function. Modern portfolio managers prioritize maximizing risk-adjusted returns over solely maximizing the returns and minimizing the risk. Aboussalah and Lee [41] employed the regularized Sharpe ratio, whereas Sood et al. [42] utilized the Differential Sharpe ratio (DSR) as a reward function in online learning settings.

Every reward function is significant in the portfolio optimization problem. The choice of an adequate reward function depends on the problem's nature and investors' risk appetite. However, choosing an optimal reward function poses a challenge for a risk-averse investor seeking to maximize returns while minimizing risk or having multiple investment goals. One possible approach is to formulate a multi-objective problem. However, there are inherent challenges in solving this problem within a RL framework. To address the abovementioned challenges, we propose a novel risk-adjusted deep reinforcement learning (RA-DRL) methodology that integrates DRL agents trained with three different reward functions: (1) log return to maximize the returns, (2) DSR to enhance the risk-adjusted returns, and (3) maximum drawdown (MDD) to manage the downside risks. Each reward function captures the distinct aspects of the portfolio optimization and provides the relationship between risk and return as a trade-off. RA-DRL combines the actions of these three agents by using a convolutional neural network (CNN) and leverages their individual strengths aiming to maximize the portfolio returns with considerable risk. As CNN training requires labeled data, we suggested a new approach that uses historical data to pre-train our RA-DRL model in a supervised setting. By using supervised learning, we can explicitly define the desired outputs based

on expert demonstrations derived from historical data. This setting allows the model to learn how to optimally weigh the contributions of each agent, leading to a more balanced and risk-adjusted decision-making process. By structuring the training this way, the approach enhances learning efficiency and ensures better generalization to unseen market conditions. Integrating supervised learning with RL provides a novel and practical solution to portfolio optimization, making the model more robust and effective for real-world financial markets. The efficacy of the proposed approach is tested across four global markets: Sensex, Dow, TWSE, and IBEX, based on the most widely used measures. The empirical study showed the superiority of our proposed method when compared against several benchmarks.

The main contributions of our paper are highlighted as follows:

1. We propose a risk-adjusted DRL (RA-DRL) framework that produces risk-adjusted returns for portfolio optimization while addressing multiple diverse objectives simultaneously.
2. The methodology offers an effective way of utilizing a CNN to extract meaningful representations from diverse actions by optimally weighing them to aggregate into a unified action.
3. We also propose a novel approach for integrating a supervised setting within a Deep RL framework for portfolio optimization by leveraging historical data.

1.1 Organization of the Paper

The organization of the paper is as follows: Sect. 2 provides a brief review of the existing literature on portfolio optimization using DRL. Section 3 outlines the problem formulation and details the proposed RA-DRL methodology. Section 4 covers the experimental setup, data description, and comparison of the proposed methodology with base DRL models and benchmarks. Finally, Sect. 5 offers concluding remarks and discusses the future implications of the presented work.

2 Related Work

This section reviews the existing applications of DRL in portfolio optimization-based solutions. The applications of RL in asset allocation may typically be divided into two parts: (1) prediction-based and (2) portfolio optimization-based solutions. Predictive models seek to predict short-term variations in asset prices to take advantage of them when making stock selection decisions. In contrast, portfolio optimization models prefer to concentrate on choosing the best portfolio, as initially encouraged by Modern Portfolio Theory (MPT) [1]. The Efficient Market Hypothesis [43] states that it is impossible to constantly beat the market by timing the stock prices. Hence, researchers often seek to identify a group of assets that will likely outperform the market on a near-term basis rather than directly forecasting market changes. The portfolio optimization problem was previously approached using conventional dynamic programming methods, with various degrees of success before the introduction of RL. Brennan et al. [44] utilized the risk-free rate, the long-term bond rate, and the dividend yield of the stocks to formulate the portfolio management problem as a Markov decision process (MDP). Using the empirical data, they determined an ideal method to distribute wealth among the asset classes, and the out-of-sample simulations supported this estimate. According to Neuneier [45], dynamic programming performed as well as Q-learning in the context of the German stock market, where it was applied for asset allocation, leveraging the artificial exchange rate as a state variable. Moody et al. [46] employed recurrent reinforcement learning (RRL) to optimize the portfolio's risk-adjusted returns and extended their work to a direct policy optimization approach [47], nowadays widely known as Direct Reinforcement. The direct reinforcement approach stands apart from dynamic programming and RL algorithms as it does not involve estimating a value function.

Deep reinforcement learning (DRL) has emerged as a powerful paradigm for dynamic decision-making in complex environments for a wide range of applications. Over the last several years, researchers have showcased

the effectiveness of DRL algorithms for solving the portfolio optimization problem. Lu [48] employed Long Short-Term Memory (LSTM) along with the policy gradient method for designing a Forex trading model. The agent exhibited the potential to manage downside risk associated with the exchange rate volatility. A model free and policy based DRL model was effectively used for managing cryptocurrency portfolios by [49]. In this study, the author established a framework that can be adapted to accommodate different variants of Deep Neural Networks. In addition, it can linearly scale the portfolio size by employing an Ensemble of Identical Independent Evaluators (EIIE) meta-topology. Aboussalah and Lee [41] proposed a Stacked Deep Dynamic Recurrent Reinforcement Learning (SDDRRL) methodology to optimize real-time portfolios. The efficacy of the proposed technique is tested on 10 different stocks chosen from various sectors of the S&P 500.

The reward function defines the agent's goal and directs it to behave optimally. Recently researchers have designed different reward functions to fulfill their investment objectives and speed learning. In their studies, [20, 34, 36] used portfolio values or log returns as reward functions to achieve significantly higher returns. Benhamou et al. [36] formulated the traditional portfolio optimization problem into a DRL framework. They considered the portfolio's net performance as a reward function to bridge the gap between conventional portfolio optimization techniques and DRL. Jiang and Liang [20] trained a model less CNN in a RL manner and used portfolio value as a reward signal. By integrating a DL approach with modern portfolio theory, [34] proposed a DRL mechanism. This proposed mechanism adopted the Tucker decomposition to deal with the multi-modal problem and continuously adjusted the weights by considering the market movements. One major issue with focusing solely on portfolio return as a reward is that it does not consider the risk when making investment decisions. Zhang et al. [50] employed direct policy gradient optimization and introduced a novel cost-sensitive reward function. The authors try to maximize cumulative returns while effectively managing associated risk costs. A novel DRL framework, DeepTrader, proposed by [40] to optimize risk-return balanced policies. The proposed strategy leverages the negative maximum draw-down as a reward function and an asset scoring unit that hierarchically represents the temporal and spatial connections between the assets. Wu et al. [25] recommended a probability-based mechanism for developing a risk prediction model. The reward function was adjusted to make the agent risk-aware by taking the downside risk measure into account. These studies [25, 40, 50] utilized the risk-sensitive reward functions, focus primarily on mitigating losses. While this is beneficial in volatile or bearish market conditions. It may lead to overly conservative investment strategies that fail to capture high returns during bullish trends. Recently, risk-adjusted returns have gained immense popularity among researchers due to their power to maximize returns while simultaneously reducing risk. Almahdi and Yang [39] proposed a technique that yields asset allocation weights and buy/sell signals under a coherent risk-adjusted reward function. Yoo et al. [30] suggested a safety framework for asset allocation that incorporates various dynamic asset allocation algorithms to produce risk-adjusted returns. Additionally, a new reward function was employed by combining the three risk-adjusted rewards: Sortino ratio, Sharpe ratio, and Calmar ratio, to optimize the PPO agent's learning process. Sood et al. [42] provided an exhaustive comparison of MVO with the model-free DRL for optimum portfolio allocation. They implemented DRL for portfolio optimization by employing a risk-adjusted reward, namely the differential shape ratio (DSR), in online learning settings, mentioning the same adjustments required to make MVO functional.

Although the DRL has gained popularity among researchers for portfolio optimization, most existing studies focus on single reward functions and do not adequately consider the trade-off between return and risk. Moreover, current studies do not effectively integrate multiple reward functions into a cohesive portfolio strategy. Thus, it encourages to develop a strategy that encapsulates multiple facets of the stock market, integrating return maximization, risk-adjusted performance, and downside risk management to cater to the diverse objectives of investors.

3 Methodology

The portfolio optimization problem refers to dynamically redistributing capital into a pool of various financial assets to minimize risk while maximizing the long-term cumulative returns. Due to the sequential decision making

nature of the portfolio optimization problem, Markov decision process (MDP) provides an ideal framework for addressing such problems [44, 51]. An MDP is defined as a tuple (S, A, P, R, γ), where S is the state space, A is the action space, P is the transition probability, R is the reward function, and γ is the discount factor lie in [0, 1]. A policy π is a solution to this MDP, which specifies the actions $\pi(s)$ performed by an agent in state s. The goal is to determine a policy, π, over a possibly infinite time horizon that maximizes the expected discounted cumulative rewards, i.e. $\mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t R(s_t, a_t)\right]$ by interacting with the environment.

3.1 Environment

In an RL setting, the environment represents the surroundings with which the agent interacts and learns. It consists of the state space, the action space, and the reward function.

3.1.1 State Space

A state of the environment consists of the relevant information at the time step t that encompasses the market conditions, technical indicators, and other relevant factors. Specifically, the state space s_t includes the covariance matrix of the closing price of the assets and eight technical indicators (30 and 60 day simple moving averages (SMA), moving average convergence divergence (MACD), upper and lower Bollinger bands, Relative Strength Index (RSI), Commodity Channel Index (CCI), and Average Directional Index (ADX)) [52] corresponding to each asset.

3.1.2 Actions Space

The actions are represented as an n-dimensional vector $w = [w_1, w_2, \ldots, w_n]$, with each component w_i signifying the allocation of weight to asset i, ensuring that $\sum_{i=1}^{n} w_i = 1$, with $0 \leq w_i \leq 1$. To enforce these constraints, one can utilize the softmax activation function on an agent's continuous actions of dimension n, where n denotes the number of assets in the portfolio.

3.1.3 Reward

The reward function incentivizes the agent to learn and refine its policy. In this study, we utilize three diverse reward functions to harness their respective advantages. A detailed description of these reward functions is as follows:

1. **Portfolio logarithmic returns** The portfolio log return [20, 34, 37, 53], denoted as $r(t)$, is defined as the logarithm of the ratio of the portfolio value at time t, P_t to the portfolio value at time $t-1$, P_{t-1} i.e.

$$r(t) = log\left(\frac{P_t}{P_{t-1}}\right)$$

The logarithmic return measures the relative change in the portfolio value and practitioners try to maximize it. The logarithmic return encourages the agent to maximize portfolio growth over time.

2. **Differential Sharpe ratio (DSR)** Most researchers and portfolio managers aim to maximize risk-adjusted returns rather than only returns. The Sharpe ratio (SR), the ratio of the portfolio return to its standard deviation, is the most popular metric for this. However, it is not ideal for online learning environments because it is defined over a specific time period. The DSR [46] is employed to counter this issue since it accesses the risk-adjusted returns at each time step t. DSR_t is derived by expanding the SR_t to the first order with respect to the adaptation

rate (η): $SR_t \approx SR_{t-1} + \eta DSR_t|_{\eta=0} + O(\eta^2)$ and formulated as follows:

$$DSR_t \equiv \frac{d}{d\eta} SR_t = \frac{Y_{t-1}\Delta X_t - \frac{1}{2} X_{t-1}\Delta Y_t}{(Y_{t-1} - X_{t-1}^2)^{\frac{3}{2}}}$$

where X_t and Y_t represent the exponential moving averages of the 1^{st} and 2^{nd} order moments of the return r_t, respectively, with

$$X_t = X_{t-1} + \eta\Delta X_t = X_{t-1} + \eta(r_t - X_{t-1})$$
$$Y_t = Y_{t-1} + \eta\Delta Y_t = Y_{t-1} + \eta(r_t^2 - Y_{t-1}).$$

The initial values for the quantities X_t and Y_t are $X_0 = Y_0 = 0$. As there are nearly 252 trading days in a year, the adaptation rate is $\eta \approx \frac{1}{252}$. DSR ensures that the agent prioritizes risk-adjusted returns instead of merely focusing on high returns, which could be volatile and unsustainable.

3. **Maximum drawdown (MDD)** The MDD [54] offers an ideal way to capture the risk throughout the entire investment period. It quantifies the largest decline from a portfolio's peak to its lowest point, providing a clear measure of downside risk. For time τ, MDD can be defined as

$$MDD = \max_{t:\tau>t} \left[\frac{P_t - P_\tau}{P_t} \right]$$

where P_t denotes the portfolio value at time t. We set the current portfolio value as the trough for our online learning setting. This approach guides the agent to take actions that push the current portfolio value to the peak. MDD protects against severe portfolio losses by discouraging allocations that lead to large drawdowns.

By integrating these metrics, the reward function provides a holistic evaluation of the portfolio's performance. It ensuring that the agent does not overly favor the high-risk assets or ignore the downside risk. This reward function design ensures a comprehensive approach to portfolio optimization, capturing the strengths of each metric while mitigating their individual weaknesses. The integration of LR, DSR, and MDD enables the RL agent to develop a more stable, risk-aware, and return-optimized policy.

3.2 Agent: Proximal Policy Optimization (PPO)

Reinforcement learning (RL) algorithms are broadly classified into two distinct categories known as model-based and model-free algorithms. The model-based algorithms perform an action that aims to maximize the reward irrespective of the consequences caused by actions. On the other hand, model-free algorithms, such as Policy Gradient and Q-learning, attempt to learn the outcomes of their actions through experiential learning. The policy gradient algorithms can be highly sensitive to perturbations as minor variations in the network parameters may lead to significant policy changes, resulting in a substantial decline in the agent's performance. A stochastic policy gradient algorithm, PPO [33], is employed to overcome this limitation by using a clipped objective function and restricting the updates in the policy network. PPO incorporates two different networks, Actor and Critic. The actor-network determines the optimal policy and generates the actions based on this policy. The critic network estimates the value function and evaluates the efficiency of the actions, which enables the actor to get feedback and enhance its decision-making process. PPO strengthens the training stability of the agent and prevents excessively large policy updates by clipping the probability ratio, which indicates the difference between the current policy and the previous one. The objective function of PPO is defined as:

$$L(\theta) = \hat{\mathbb{E}}_t \left[\min \left\{ \frac{\pi_\theta(a_t|s_t)}{\pi_{\theta_{old}}(a_t|s_t)} \cdot \hat{A}_t, \text{clip}\left(\frac{\pi_\theta(a_t|s_t)}{\pi_{\theta_{old}}(a_t|s_t)}, 1-\epsilon, 1+\epsilon \right) \cdot \hat{A}_t \right\} \right]$$

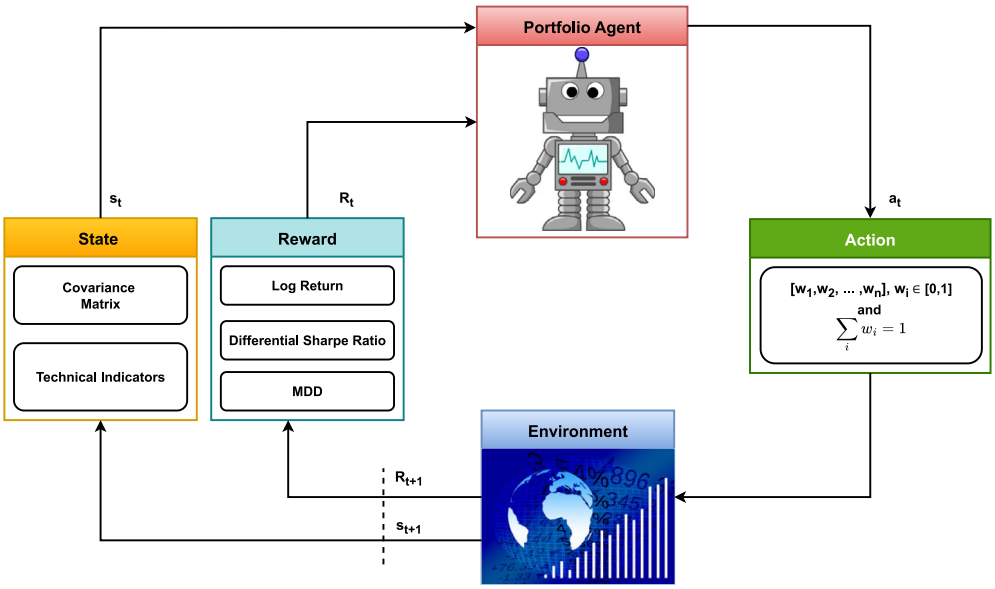

Fig. 1 Deep reinforcement learning framework for portfolio optimization

where $\frac{\pi_\theta(a_t|s_t)}{\pi_{\theta_{old}}(a_t|s_t)}$ represents the likelihood ratio at time t. This ratio is clipped within the range $[1-\epsilon, 1+\epsilon]$, thereby discouraging significant deviations of the current policy from the old one and $\hat{A}_t$ is the advantage function, defined as $\hat{A}_\pi(s_t, a_t) = Q_\pi(s_t, a_t) - V_\pi(s_t)$, compares the effectiveness of action a_t to the average of other actions taken at that particular state s_t.

3.3 Proposed Methodology

This study proposes RA-DRL, a risk-adjusted DRL framework for optimizing portfolio weights under risk uncertainties. It leverages three DRL models trained via distinct reward functions and provides a unified framework to serve desirable goals. In RL, the reward function defines the agent's objective and guides learning. A risk-averse investor could have multiple goals, such as maximizing the return, minimizing the risk, or both simultaneously. Our study considers three investment goals: maximize the return, minimize the risk, and maximize the risk-adjusted return. In the RL setting, these three objectives are represented by three distinct reward functions: Log return to maximize the returns, DSR to maximize the risk-adjusted returns, and MDD to minimize the risk. Three different PPO agents are then trained using these reward functions. Fig. 1 depicts the DRL framework adopted for portfolio optimization.

Three trained agents correspond to three distinct actions for each time step. Subsequently, these actions are aggregated into a unified volume. A $2D$ CNN layer employing a kernel size of $(1, 3)$ is then utilized to extract meaningful representations from these stacked actions. This choice of kernel size is motivated by the objective of assigning weights to actions associated with each of the three DRL agents, enabling a focused representation extraction process. These extracted representations are then fed to fully connected layers to produce the final action or weights for the portfolio. Fig. 2 illustrates the complete flow of the RA-DRL methodology. These weights are then utilized to compute the portfolio's return. The network parameters are adjusted to minimize the difference between this portfolio return and the ground truth. Ground truth plays a vital role in assessing model performance and modifying parameters by providing essential feedback in the form of gradients. In this study, we propose a novel approach that uses a supervised setting within a DRL framework. Due to the availability of an ample amount

 https://doi.org/10.1007/s44196-025-00875-8

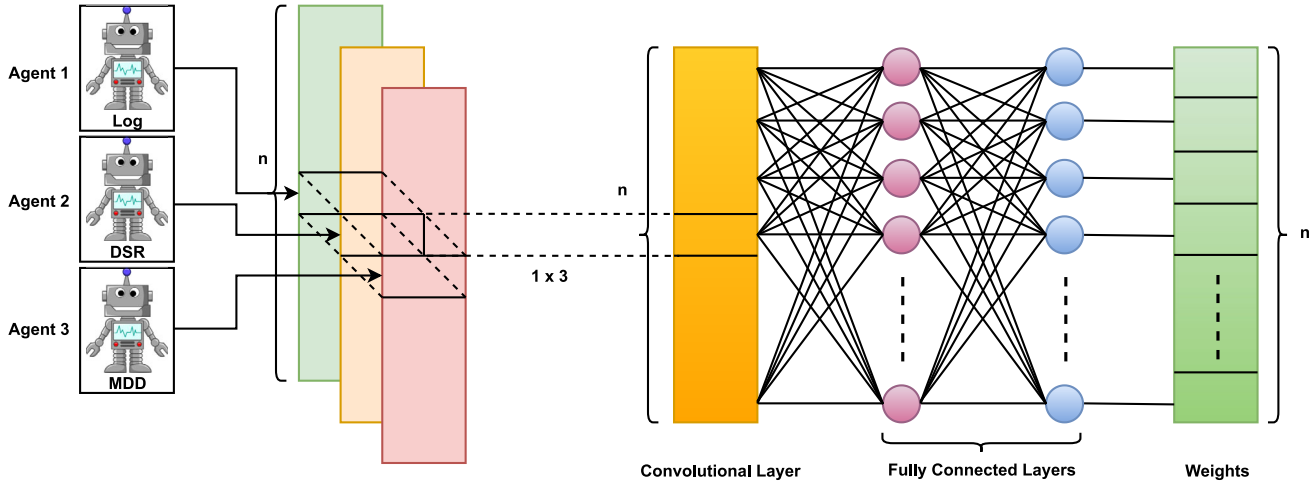

Fig. 2 The overview of the RA-DRL methodology, where the actions generated by the base DRL agents are combined using a CNN layer. Then, fully connected layers are utilized to produce the risk-adjusted portfolio weights

of historical data during the training phase, we compute the weights of the portfolio assets using the following:

$$w_{i,t} = \frac{e^{\rho_{i,t} \times c}}{\sum_i e^{\rho_{i,t} \times c}} \tag{1}$$

where c is a constant that takes values from 1 to 5 and $\rho_{i,t}$ is the percentage change in the price of i^{th} stock at time t. The choice of this formulation for calculating portfolio weights is motivated from [55]. These weights are required to calculate the portfolio return using historical returns, which acts as the ground truth for training our proposed methodology.

4 Experiments

In this section, we provide a detailed experimental illustration of the experiments to exhibit the superiority of the proposed RA-DRL methodology against the base DRL agents and benchmarks.

4.1 Data Description

To evaluate the effectiveness of our proposed methodology for solving the portfolio optimization problem, we selected the following four distinct stock markets: (1) Sensex from India, (2) Dow from the USA, (3) TWSE from Taiwan, and 4. IBEX from Spain. For each market instance, we are taking daily data of each market's top 30 stocks according to the market capitalization from 01/01/2011 to 31/03/2024. During this period, data for only 29 and 28 stocks out of 30 are available for Dow and IBEX, respectively. For Sensex and TWSE, the data for all 30 stocks is available. The dataset consists of Open, High, Low, and Close prices collected from Yahoo Finance. In-sample data from 01/01/2011 to 31/12/2020 is utilized for training the PPO agents, and their trading performance is assessed on out-of-sample data spanning from 01/01/2021 to 31/03/2024.

4.2 Experimental Setup

All the empirical studies were conducted in the same environment settings. For tuning the model's hyperparameters, Bayesian optimization [56] is utilized on 10% of the training data. For this purpose, we leveraged a Python module

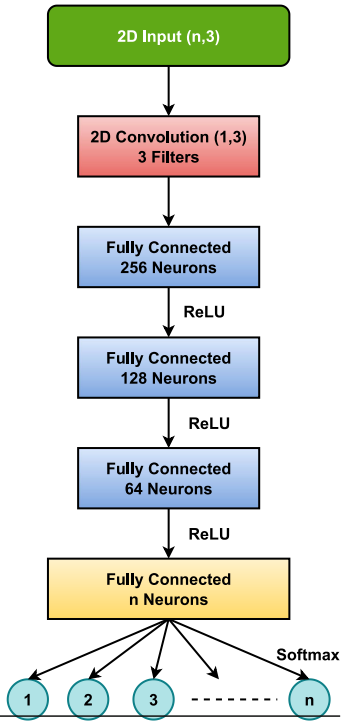

Fig. 3 Visual representation of the hybrid CNN-MLP architecture (n denotes the number of stocks in the portfolio)

Table 1 Parameters setting for experiment

Parameter	Range
Hidden layers	[1, 8]
Hidden layers dim	[2, 512]
Learning rate	$[10^{-8}, 10^{-1}]$
Discount factor (γ)	[0, 1]
Activation function	[Relu, Sigmoid, Tanh]
Dropout rate	[0, 0.5]
Entropy coefficient	[0.01, 0.1]
Value function coefficient	[0.5, 1]
PPO epochs	[5, 50]
ϵ-clip	0.2
No of episodes	500

known as Hyperopt. The range of hyperparameters picked based on empirical studies [33, 57, 58] for all three agents is reported in Table 1. A visual representation of the hybrid CNN-MLP architecture employed for the supervised training of the RA-DRL framework is presented in Fig. 3. The transaction cost at time step t can be calculated as $\delta_t = \beta \times P_t \times |w_t - w_{t-1}|$. Based on the historical studies [59, 60], the transaction rate β is set to 0.05% for all the experiments.

4.3 Performance Metrics

We have assessed the effectiveness of our proposed approach using a range of performance metrics. Specifically, we are employing eight widely recognized risk and return measures for this evaluation. Annual returns (AR) and Cumulative returns (CR) offer valuable insights into a portfolio's long-term performance. However, they do not account for risk. One popular approach involves utilizing the annualized standard deviation of daily returns to gauge volatility risk. Additionally, we incorporate the Sharpe ratio (SR), Omega ratio (OR), Calmar ratio (CAR), and Sortino ratio (SOR) to assess risk-adjusted returns. We have also applied a stability metric to review the stability of the investment strategy. These performance metrics can be defined as follows:

- **Cumulative return (CR)** Cumulative return is the overall returns generated at the end of the trading period and is calculated by dividing the portfolio's cumulative wealth by its initial wealth.
- **Annual return (AR)** Annual return, expressed as a percentage of the initial investment amount, measures the average return accumulated in a year.
- **Sharpe ratio (SR)** [61] SR measures the excess return that a trader makes per unit risk and is defined as

$$SR = \frac{r_p - r_f}{\sigma_p}$$

where r_p and r_f are the portfolio return and risk-free rate, respectively. σ_p is the standard deviation of portfolio return.
- **Calmar ratio (CAR)** [62] CAR measures the excess return that a trader makes relative to the drawdown risk of the portfolio. MDD is the measure of the largest drop in the portfolio's value from its peak to its trough. CAR signifies an investment's ability to recover from losses and is defined as

$$CAR = \frac{r_p - r_f}{MDD}$$

where MDD is the maximum draw-down.
- **Sortino ratio (SOR)** [63] While the Sharpe ratio considers both upside and downside risks equally, SOR measures only the excess return that a trader makes per unit of downside risk. A downside risk is a potential loss of money on the total investment. The Sortio ratio is defined as:

$$SOR = \frac{r_p - r_f}{DR}$$

where DR is the standard deviation of downside risks.
- **Omega ratio (OR)** [64] OR encompasses all information associated with risk and return and measures the probability-weighted ratio of profits over losses at a particular level of expected return. The Omega ratio is defined as

$$OR = \frac{\int_{\theta}^{b}(1 - F(r))dr}{\int_{a}^{\theta} F(r)dr}$$

where θ is the investor-preferred return, commonly known as a threshold, and F is the cumulative distribution function of the return r.
- **Annual volatility (AV)** Portfolio volatility serves as a metric for portfolio risk, indicating the degree to which a portfolio's returns deviate from its average return. The annualized volatility is determined by multiplying the standard deviation by the square root of the number of periods in a year, i.e., $\sigma_p \times \sqrt{T}$. Since there are approximately 252 trading days in a year, we set T equal to 252.
- **Stability** Stability denotes the capability of maintaining consistent trends or patterns over an extended period. Calculating the R-squared value of a linear regression fit to the cumulative log returns is a common way to measure stability. The stability of a trading strategy is essential for investors seeking consistent and reliable returns with continuous improvement in their investment approach.

4.4 Benchmarks Comparisons

To establish the proposed strategy's efficacy, we compare it against several standard benchmarks based on the abovementioned performance metrics. A brief description of these benchmarks is as follows:

- **Market Index** A market index is the price-weighted index comprising the prominent stocks of the exchange, offering a reflection of the broader market. The BSESN, DJIA, IBEX35, and TWII indices served as benchmarks for the Sensex, Dow 30, IBEX, and TWSE stock markets, respectively.
- **Mean-variance optimization (MVO)** [1] MVO provides a framework for determining the optimal allocation in portfolio optimization and seeks to minimize portfolio risk, or more specifically, portfolio volatility while maximizing portfolio returns.
- $1/n$ **investment strategy** This strategy split the total wealth uniformly among the all stocks in the portfolio.
- **Single objective reward** The problem of multiple objectives is approached as a single objective problem through a weighted sum approach. Here, the weights assigned to each objective are considered hyperparameters and are optimized effectively. This single-objective formulation also serves as a benchmark strategy for our comparative analysis.

4.5 Results and Discussions

In our study, we backtested the proposed RA-DRL methodology for the trading period from January 1, 2021, to March 31, 2024. An initial capital of 10,00,000 is allocated to the agent at the beginning of the trading period. Our analysis focuses on two main aspects of the results. Firstly, the RA-DRL is assessed against the base model with different objectives used by recent studies [34, 40, 42], and subsequently, it is evaluated against the benchmarks. All the simulations are conducted in similar environment settings to maintain uniformity among results. The performance comparison of all models, based on CR, SR, AV, OR, Stability, and AR, is reported in Table 2. The best results are highlighted in bold.

For the Sensex index, our proposed methodology produced the highest cumulative and annualized returns of 124.83% and 29.03%, respectively, among the base models and the benchmarks. When the market index gave a cumulative return of 52.49% during the trading period, RA-DRL achieved almost 1.4 times more returns than the index. Regarding returns, our method was followed by single objective, Log, DSR, and $1/n$ strategy, but a significant margin trails them. The risk of a portfolio is measured by the volatility of the returns. The MVO model achieves the least volatility of 12.17%, followed by RA-DRL and MDD. However, the stability of RA-DRL is significantly better than other models. Additionally, the results are noteworthy when risk-adjusted returns are taken into account. RA-DRL maintains the highest Sharpe ratio 1.69 and Omega ratio 1.33 among the models and benchmarks, indicating its ability to generate higher risk-free returns. The benchmarks single objective and MVO surpassed the base DRL models in terms of these ratios, but our proposed RA-DRL outperformed them substantially. This illustrates the effectiveness of our model in achieving risk-adjusted returns. Overall, out of six measures considered in Table 2, RA-DRL exhibits superior performance in five, except for the volatility. The rise in volatility stems from the proposed methodology's objective of achieving high returns while maintaining sustainable risk levels. Also, it is interesting to note that among the base models, the Log agent demonstrates strong performance in return measures, MDD in risk measures, and DSR in risk-adjusted measures. However, our proposed RA-DRL outperforms these base DRL agents across all measures.

Figure 4 presents the cumulative wealth plot of all models for the trading period for the Sensex data. During the first year of trading, the market showed a bullish trend, and all the models accumulated wealth. The proposed RA-DRL agent is achieving the highest wealth, followed by Log which solely focuses on maximizing returns and single objective agent. For the second year, the market showed a flat trend and sideways movement. As there was no significant market movement, most of the models started to lose wealth. However, RA-DRL continued its profitability during this cycle. This demonstrates its efficacy in capturing short-term trends while mitigating the risk. The markets went bullish again in the last trading cycle, and all models accumulated wealth again. During this period, RA-DRL consistently maintained profitability while simultaneously keeping the volatility in check. Overall, our proposed methodology significantly surpassed the benchmarks and other DRL models, accumulating the highest cumulative wealth. Moreover, the greater stability exhibited by the RA-DRL approach indicates consistency throughout the entire trading period.

 https://doi.org/10.1007/s44196-025-00875-8

Table 2 Performance indicators on Sensex, Dow, TWSE, and IBEX for trading period

Model/benchmark	CR	SR	AV	OR	Stability	AR
Sensex 30						
RA-DRL	**124.8329%**	**1.692785**	14.2016%	**1.335049**	**0.936809**	**29.0320%**
Log [34]	114.0352%	1.612136	15.5076%	1.305922	0.907859	27.0495%
DSR [42]	106.6776%	1.625449	14.9273%	1.319015	0.867024	25.6589%
MDD [40]	101.1960%	1.606143	14.3405%	1.316238	0.931517	24.6008%
Single objective	116.5389%	1.654626	15.4154%	1.327764	0.904127	27.5152%
MVO	88.5056%	1.625633	**12.1653%**	1.317586	0.911773	22.0728%
$1/n$	105.4553%	1.587932	14.5486%	1.287212	0.851483	23.4455%
Sensex	52.4919%	0.984996	14.6349%	1.181017	0.798164	14.2529%
Dow 30						
RA-DRL	**50.7842%**	**1.00689**	13.5400%	**1.186718**	0.393546	**13.7842%**
Log [34]	40.6031%	0.779999	14.9622%	1.141663	0.379436	11.1262%
DSR [42]	39.6779%	0.759364	15.1333%	1.136967	0.223607	10.8993%
MDD [40]	37.4505%	0.703641	13.4019%	1.126203	0.185654	10.3488%
Single objective	41.7906%	0.815846	15.4582%	1.150679	0.380873	10.6808%
MVO	27.2869%	0.634583	**12.1335%**	1.099574	**0.535129**	7.7554%
$1/n$	32.8548%	0.702523	13.4785%	1.124578	0.226648	8.9745%
DJI	31.5518%	0.654518	14.6425%	1.117923	0.117474	8.8719%
TWSE 30						
RA-DRL	**96.7025%**	**1.660754**	13.6612%	**1.327471**	0.806074	**24.2908%**
Log [34]	74.1795%	1.392760	13.4620%	1.268751	0.812507	19.5263%
DSR [42]	72.8329%	1.394959	13.2405%	1.268102	0.847796	13.2285%
MDD [40]	69.2874%	1.370261	12.9668%	1.260149	0.811074	18.4368%
Single objective	85.4521%	1.510955	13.7725%	1.295079	0.808369	21.9600%
MVO	51.9445%	1.147889	**8.3928%**	1.300446	**0.851280**	14.3928%
$1/n$	89.4878%	1.524856	15.4258%	1.274585	0.745487	20.1889%
TWSE	35.1933%	0.666083	16.6774%	1.120848	0.000484	10.1911%
IBEX 30						
RA-DRL	**70.5627%**	**1.066588**	16.4560%	**1.194400**	0.750988	**17.5760%**
Log [34]	62.7127%	0.974762	16.5555%	1.176544	0.748863	15.9080%
DSR [42]	62.4046%	0.980855	16.3616%	1.175784	0.777546	15.8414%
MDD [40]	58.9013%	0.939393	16.3832%	1.168096	0.761565	15.0778%
Single objective	63.0955%	0.966703	16.8121%	1.173173	**0.820348**	15.9906%
MVO	49.8123%	0.874523	**13.5484%**	1.155327	0.659099	13.0406%
$1/n$	59.4578%	0.647852	16.4125%	1.171566	0.704584	15.1245%
IBEX	37.1901%	0.669325	16.3660%	1.119047	0.336418	10.0761%

Best results are highlighted in bold

Our proposed RA-DRL agent outperformed the other models by a substantial margin for the Dow index, generating cumulative and annualized returns of 50.78% and 13.78%, respectively. Like the Sensex index, RA-DRL surpassed the Dow index returns by 1.6 times. The agent with the MDD reward function exhibits a minimum annual volatility of 13.40% among all the deep RL models, slightly bettering RA-DRL. Overall, the MVO model has the lowest volatility among baselines and benchmarks. Also, it shows the best stability, followed by RA-DRL. The low volatility of the MVO model is attributed to achieving low returns, even failing to beat the market index. Having the highest Sharpe and Omega ratios of 1.01 and 1.19 demonstrates that RA-DRL effectively produces

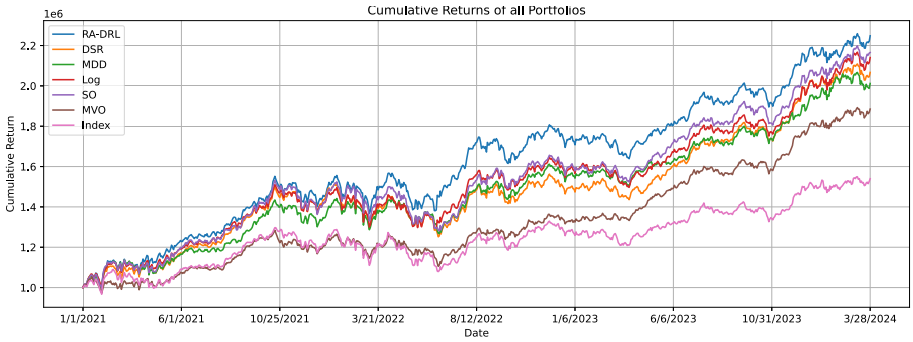

Fig. 4 Comparison of the proposed RA-DRL with base DRL agents and benchmarks based on the cumulative wealth for the trading period on Sensex

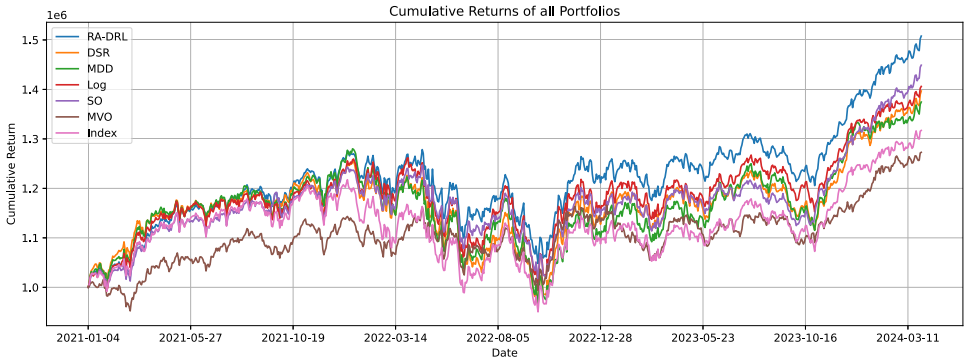

Fig. 5 Comparison of the proposed RA-DRL with base DRL agents and benchmarks based on the cumulative wealth for the trading period on Dow

risk-adjusted returns compared to all the base DRL agents and benchmarks. All these performance metrics affirm that our proposed RA-DRL model produces outstanding returns at a tolerable risk.

The cumulative wealth plot of all the models for the Dow trading data is depicted in Fig. 5. The market was bullish for the first six months of the trade, and most models accumulated wealth during this period. The MVO model, which aims to minimize the volatility while maximizing the returns, lagged significantly behind all other models due to high volatility in the market. During the next two quarters, the market was sideways, and the RA-DRL agent achieved the highest cumulative wealth, closely followed by MDD and Log agent. For the second year, the market showed a bearish trend. The dip in the stock market during 2022 was an economic event that affected the stock markets globally. The majority of the models began to lose wealth during this period. Notably, the MVO model generates superior cumulative wealth compared to the Log and DSR agents during this dip, attributed to its ability to manage risk while maximizing returns. However, RA-DRL, showcasing its risk-averse capability, lost the least amount and surpassed other models. During the last trading cycle, the market again showed an upward trend. The proposed RA-DRL strengthened its position, outperformed the other DRL models and benchmarks, and demonstrated its superiority.

A similar performance can also be observed for the TWSE and IBEX Index, where our proposed approach yields superior cumulative and annual returns. The cumulative return of RA-DRL for the TWSE Index is 96.70%, approximately 1.75 times higher than the market returns. For the IBEX index, RA-DRL generates a cumulative return of 70.56%. Additionally, Table 2 shows that the risk-adjusted returns of RA-DRL for both indices are also significantly higher than the benchmarks. In TWSE, the MVO model shows better stability by overtaking the DRL models by a slight margin. However, the DRL agents and MVO illustrate notably superior strength compared to the market index. Moreover, the DRL agents outshined their stability for the IBEX Index, outperforming MVO and the market index by a significant margin. The plots for cumulative wealth for TWSE and IBEX Indices are displayed in Figs. 6 and 7. For the TWSE Index, RA-DRL faced challenges during the initial trading phase, and the single objective agent outperformed the benchmarks, closely followed by the RA-DRL. During the last trading phase, our proposed methodology exhibited dominance over benchmarks and other DRL agents. For the IBEX index, the proposed RA-DRL shows the preeminence performance. It surpasses the DRL agents with individual

 https://doi.org/10.1007/s44196-025-00875-8

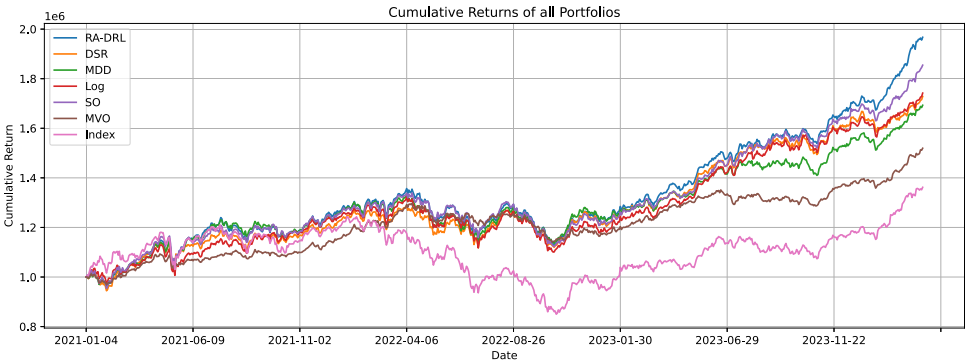

Fig. 6 Comparison of the proposed RA-DRL with base DRL agents and benchmarks based on the cumulative wealth for the trading period on TWSE

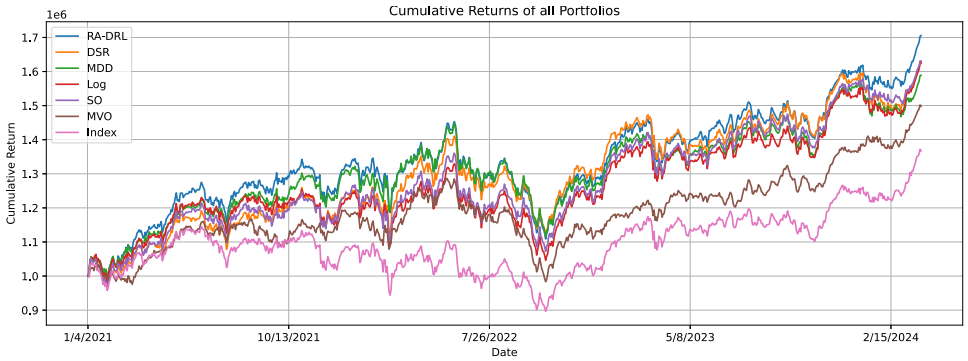

Fig. 7 Comparison of the proposed RA-DRL with base DRL agents and benchmarks based on the cumulative wealth for the trading period on IBEX

reward functions and benchmarks during the upward market trend. When the MVO model and all the DRL agents struggled during the bearish trend in the market, our proposed RA-DRL showed superior performance.

Furthermore, to demonstrate the risk-adjusted behavior of our proposed RA-DRL methodology, the Calmar ratio and Sortino ratio plots of all models for the four market instances are showcased in Figs. 8 and 9 respectively. The Sharpe ratio takes average deviation into account, whereas the Calmar ratio is more responsive to substantial losses, and the Sortino ratio considers the negative deviation, offering more risk-averse returns. It is evident from Fig. 8 that RA-DRL attains the highest Calmar ratio among all models for the Sensex, Dow, and TWSE data. However, for the IBEX market, the single-objective DRL agent emerges as the top performer, followed by RA-DRL and DSR. Our proposed method closely resembles the performance of the single objective model, and the difference between these two is minimal. A similar inference can be drawn from Fig. 9, where RA-DRL outperforms the other models by a significant margin for Sensex, Dow, and TWSE markets. For the case of the IBEX market, RA-DRL is yielding nearly comparable results to the MVO model and exceeding the other models by a considerable margin. Our proposed methodology exhibited superior performance compared to base models and benchmarks in generating risk-averse returns.

Statistical significance analysis of the proposed RA-DRL model To assess the statistical significance of performance differences of RA-DRL against the other DRL models, we conducted paired t-tests across multiple key performance indicators (CR, SR, OR, and SOR). The null hypothesis (H_0) states that there is no significant difference between RA-DRL and the models, while the alternative hypothesis (H_a) suggests that a significant difference exists. The p-values are reported in Table 3 for each comparison, highlighting that the proposed RA-DRL exhibits statistically significant differences compared to other DRL models across various metrics. A significance threshold of 0.05 is used to determine statistically meaningful differences, with p-values between 0.05 and 0.10 indicating marginal significance.

RA-DRL significantly outperforms Log, DSR, MDD, and SO models in CR, with p-values ranging from 0.0176 to 0.0466. In terms of SR, RA-DRL demonstrates significant differences against all benchmark models with p-values between 0.0041 and 0.0445. For the OR, RA-DRL maintains significant differences against Log and SO models, with p-values of 0.0234 and 0.0263, while its difference with the DSR and MDD models are marginally

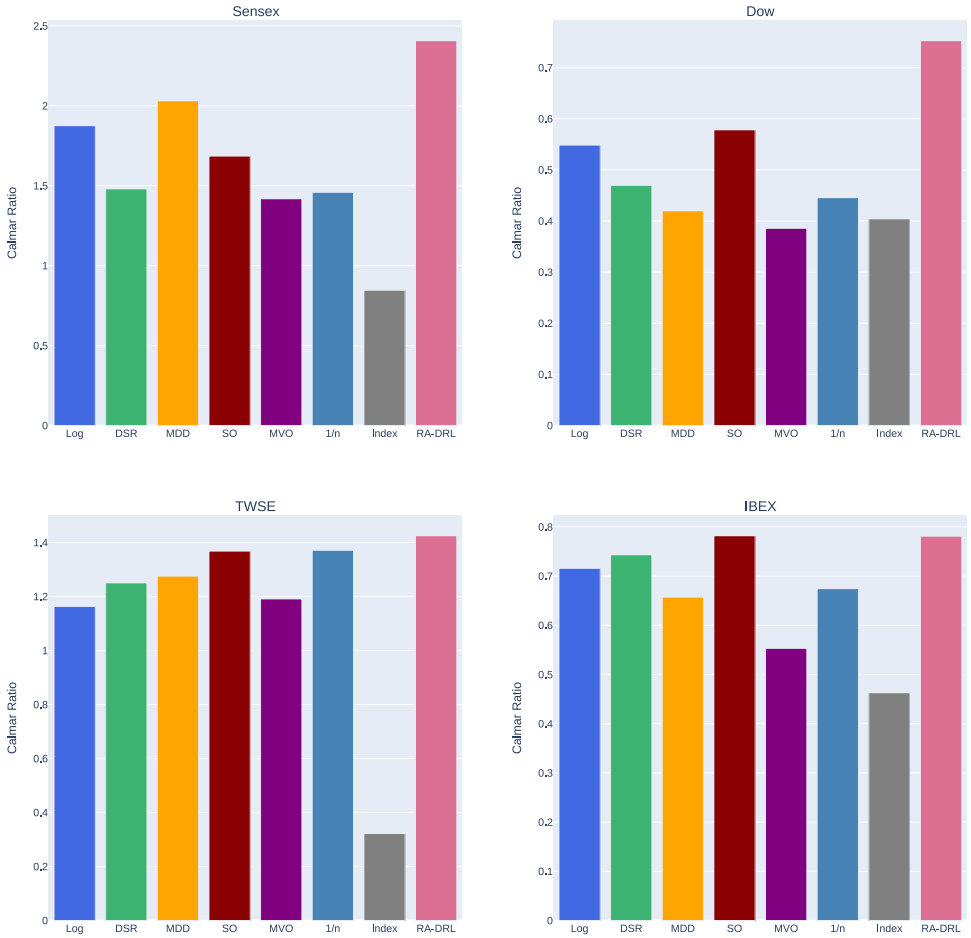

Fig. 8 Performance of Calmar ratio for the proposed RA-DRL, base DRL agents, and benchmarks on Sensex, Dow, TWSE, and IBEX

Table 3 Paired t-test results comparing RA-DRL with other DRL models across key performance indicators (CR, SR, OR, and SOR). Significant differences are identified based on a 0.05 significance threshold

Model comparison	CR	SR	OR	SOR	Result
RA-DRL vs Log	0.0262	0.0181	0.0234	0.0210	Shows significant difference
RA-DRL vs DSR	0.0466	0.0413	0.0501	0.0345	Shows significant difference
RA-DRL vs MDD	0.0465	0.0445	0.0513	0.0441	Shows significant difference
RA-DRL vs SO	0.0176	0.0041	0.0263	0.0071	Shows significant difference

significant. Similarly, For the SOR, RA-DRL shows significant differences for most models, with p-values between 0.0071 and 0.0441. Overall, these results emphasize that the proposed RA-DRL model shows significant difference across key performance measures against the other DRL models.

5 Conclusion and Future Work

This paper proposes a RA-DRL approach for portfolio optimization. RA-DRL leverages three instances of PPO agents trained via different reward functions with diverse objectives. The proposed methodology then utilizes a CNN to refine the policy learned with PPO agents, providing a robust risk-averse policy. The learned policy effectively maximizes profitability while mitigating the investment risk. A comparative analysis considering various risk and return-specific measures is conducted across four global markets against several benchmarks. The empirical study confirmed the robustness and efficacy of our proposed methodology in optimizing portfolios under diverse

 https://doi.org/10.1007/s44196-025-00875-8

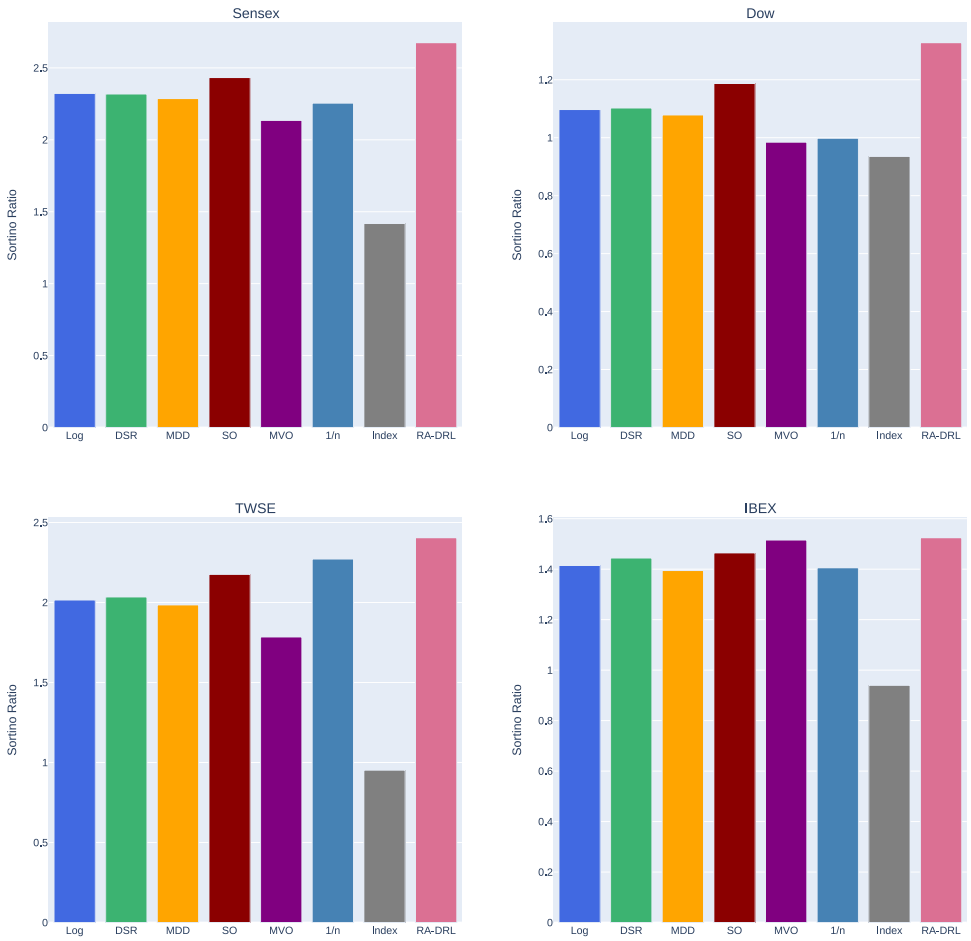

Fig. 9 Performance of Sortino ratio for the proposed RA-DRL, base DRL agents, and benchmarks on Sensex, Dow, TWSE, and IBEX

market conditions. The proposed RA-DRL model is not limited to the financial sector but can also be applied across a broad range of RL domains, addressing practical challenges associated with decision-making involving conflicting objectives.

Future applications may employ more comprehensive DRL agents to optimize portfolios. A single dynamic reward function could enhance stable learning, replacing the need for combining multiple rewards, thus suggesting a potential avenue for future research. The current portfolio optimization framework is limited to fixed number of stocks. In practice, portfolios may encompass a broader asset universe, including ETFs, bonds, and derivatives, each with distinct risk-return characteristics. Incorporating these instruments would yield a more comprehensive and realistic portfolio optimization strategy, aligning better with diverse investor objectives and dynamic market conditions.

Author Contributions Himanshu Choudhary: conceptualization, methodology, visualization, resources, investigation, data curation, software, formal analysis, writing—original draft, review and editing. Arishi Orra: conceptualization, methodology, visualization, resources, investigation, data curation, software, formal analysis, writing—review and editing. Kartik Sahoo: formal analysis, methodology, review and editing. Manoj Thakur: supervision, validation, formal analysis, resources, writing—review and editing.

Funding Open access funding provided by Siksha 'O' Anusandhan (Deemed To Be University) The financial support is received from Siksha O Anusandhan (Deemed to be) University, Bhubaneswar, 751030, Odisha, India.

Data availability statement The datasets used are publicly available. We have added references to the source of datasets in 4.1.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

Ethical and informed consent Ethical and informed consent for data used not applicable.

References

1. Markowitz, H.: Portfolio selection. J. Finance **7**(1), 77–91 (1952)
2. Levy, H., Sarnat, M.: International diversification of investment portfolios. Am. Econ. Rev. **60**(4), 668–675 (1970)
3. Davis, M.H., Norman, A.R.: Portfolio selection with transaction costs. Math. Oper. Res. **15**(4), 676–713 (1990)
4. Ehrgott, M., Klamroth, K., Schwehm, C.: An MCDM approach to portfolio optimization. Eur. J. Oper. Res. **155**(3), 752–770 (2004)
5. Anagnostopoulos, K., Mamanis, G.: A portfolio optimization model with three objectives and discrete variables. Comput. Oper. Res. **37**(7), 1285–1297 (2010)
6. Zhu, H., Wang, Y., Wang, K., Chen, Y.: Particle swarm optimization (pso) for the constrained portfolio optimization problem. Expert Syst. Appl. **38**(8), 10161–10169 (2011)
7. Sharma, A., Mehra, A.: Financial analysis based sectoral portfolio optimization under second order stochastic dominance. Ann. Oper. Res. 1–27 (2016)
8. Abi Jaber, E., Miller, E., Pham, H.: Markowitz portfolio selection for multivariate affine and quadratic Volterra models. SIAM J. Financ. Math. **12**(1), 369–409 (2021)
9. Benhamou, E., Saltiel, D., Ungari, S., Mukhopadhyay, A.: Time your hedge with deep reinforcement learning. arXiv:2009.14136 (2020)
10. Heaton, J.B., Polson, N.G., Witte, J.H.: Deep learning for finance: deep portfolios. Appl. Stoch. Models Bus. Ind. **33**(1), 3–12 (2017)
11. Ma, Y., Han, R., Wang, W.: Portfolio optimization with return prediction using deep learning and machine learning. Expert Syst. Appl. **165**, 113973 (2021)
12. Chen, W., Zhang, H., Mehlawat, M.K., Jia, L.: Mean-variance portfolio optimization using machine learning-based stock price prediction. Appl. Soft Comput. **100**, 106943 (2021)
13. Zhang, Z., Zohren, S., Stephen, R.: Deep reinforcement learning for trading. J. Financ. Data Sci. (2020)
14. Hao, J., He, F., Ma, F., Zhang, S., Zhang, X.: Machine learning vs deep learning in stock market investment: an international evidence. Ann. Oper. Res. 1–23 (2023)
15. Perrin, S., Roncalli, T.: Machine learning optimization algorithms & portfolio allocation. In: Machine Learning for Asset Management: New Developments and Financial Applications, pp. 261–328 (2020)
16. Pinelis, M., Ruppert, D.: Machine learning portfolio allocation. J. Finance Data Sci. **8**, 35–54 (2022)
17. Sutton, R.S., Barto, A.G.: Reinforcement Learning: An Introduction (2018)
18. Betancourt, C., Chen, W.-H.: Deep reinforcement learning for portfolio management of markets with a dynamic number of assets. Expert Syst. Appl. **164**, 114002 (2021)
19. Jeong, G., Kim, H.Y.: Improving financial trading decisions using deep q-learning: Predicting the number of shares, action strategies, and transfer learning. Expert Syst. Appl. **117**, 125–138 (2019)
20. Jiang, Z., Liang, J.: Cryptocurrency portfolio management with deep reinforcement learning. In: 2017 Intelligent Systems Conference (IntelliSys), pp. 905–913 (2017). IEEE

21. Liang, Z., Chen, H., Zhu, J., Jiang, K., Li, Y.: Adversarial deep reinforcement learning in portfolio management. arXiv:1808.09940 (2018)
22. Si, W., Li, J., Ding, P., Rao, R.: A multi-objective deep reinforcement learning approach for stock index future's intraday trading. In: 2017 10th International Symposium on Computational Intelligence and Design (ISCID), vol. 2, pp. 431–436 (2017). IEEE
23. Lim, Q.Y.E., Cao, Q., Quek, C.: Dynamic portfolio rebalancing through reinforcement learning. Neural Comput. Appl. **34**(9), 7125–7139 (2022)
24. Song, Z., Wang, Y., Qian, P., Song, S., Coenen, F., Jiang, Z., Su, J.: From deterministic to stochastic: an interpretable stochastic model-free reinforcement learning framework for portfolio optimization. Appl. Intell. **53**(12), 15188–15203 (2023)
25. Wu, J.M.-T., Lin, S.-H., Syu, J.-H., Wu, M.-E.: Embedded draw-down constraint reward function for deep reinforcement learning. Appl. Soft Comput. **125**, 109150 (2022)
26. Wu, M.-E., Syu, J.-H., Lin, J.C.-W., Ho, J.-M.: Portfolio management system in equity market neutral using reinforcement learning. Appl. Intell. **51**(11), 8119–8131 (2021)
27. Day, M.-Y., Yang, C.-Y., Ni, Y.: Portfolio dynamic trading strategies using deep reinforcement learning. Soft Comput. 1–16 (2023)
28. Yu, X., Wu, W., Liao, X., Han, Y.: Dynamic stock-decision ensemble strategy based on deep reinforcement learning. Appl. Intell. **53**(2), 2452–2470 (2023)
29. Mnih, V., Kavukcuoglu, K., Silver, D., Graves, A., Antonoglou, I., Wierstra, D., Riedmiller, M.: Playing atari with deep reinforcement learning. arXiv:1312.5602 (2013)
30. Yoo, S.J., Gu, Y.H.: Safety aarl: weight adjustment for reinforcement-learning-based safety dynamic asset allocation strategies. Expert Syst. Appl. **227**, 120297 (2023)
31. Sun, Q., Wei, X., Yang, X.: Graphsage with deep reinforcement learning for financial portfolio optimization. Expert Syst. Appl. **238**, 122027 (2024)
32. Winkel, D., Strauß, N., Schubert, M., Seidl, T.: Risk-aware reinforcement learning for multi-period portfolio selection. In: Joint European Conference on Machine Learning and Knowledge Discovery in Databases, pp. 185–200 (2022). Springer
33. Schulman, J., Wolski, F., Dhariwal, P., Radford, A., Klimov, O.: Proximal policy optimization algorithms. arXiv:1707.06347 (2017)
34. Jang, J., Seong, N.: Deep reinforcement learning for stock portfolio optimization by connecting with modern portfolio theory. Expert Syst. Appl. **218**, 119556 (2023)
35. Lillicrap, T.P., Hunt, J.J., Pritzel, A., Heess, N., Erez, T., Tassa, Y., Silver, D., Wierstra, D.: Continuous control with deep reinforcement learning. arXiv:1509.02971 (2015)
36. Benhamou, E., Saltiel, D., Ungari, S., Mukhopadhyay, A.: Bridging the gap between markowitz planning and deep reinforcement learning. arXiv:2010.09108 (2020)
37. Yu, P., Lee, J.S., Kulyatin, I., Shi, Z., Dasgupta, S.: Model-based deep reinforcement learning for financial portfolio optimization. In: RWSDM Workshop, ICML, vol. 1, p. 2019 (2019)
38. Orra, A., Bhambu, A., Choudhary, H., Thakur, M.: Dynamic reinforced ensemble using bayesian optimization for stock trading. In: Proceedings of the 5th ACM International Conference on AI in Finance, pp. 361–369 (2024)
39. Almahdi, S., Yang, S.Y.: An adaptive portfolio trading system: a risk-return portfolio optimization using recurrent reinforcement learning with expected maximum drawdown. Expert Syst. Appl. **87**, 267–279 (2017)
40. Wang, Z., Huang, B., Tu, S., Zhang, K., Xu, L.: Deeptrader: a deep reinforcement learning approach for risk-return balanced portfolio management with market conditions embedding. Proc. AAAI Conf. Artif. Intell. **35**, 643–650 (2021)
41. Aboussalah, A.M., Lee, C.-G.: Continuous control with stacked deep dynamic recurrent reinforcement learning for portfolio optimization. Expert Syst. Appl. **140**, 112891 (2020)
42. Sood, S., Papasotiriou, K., Vaiciulis, M., Balch, T.: Deep reinforcement learning for optimal portfolio allocation: a comparative study with mean-variance optimization. FinPlan **2023**, 21 (2023)
43. Malkiel, B.G.: The efficient market hypothesis and its critics. J. Econ. Perspect. **17**(1), 59–82 (2003)
44. Brennan, M.J., Schwartz, E.S., Lagnado, R.: Strategic asset allocation. J. Econ. Dyn. Control **21**(8–9), 1377–1403 (1997)
45. Neuneier, R.: Optimal asset allocation using adaptive dynamic programming. Adv. Neural Inf. Process. Syst. **8** (1995)
46. Moody, J., Wu, L., Liao, Y., Saffell, M.: Performance functions and reinforcement learning for trading systems and portfolios. J. Forecast. **17**(5–6), 441–470 (1998)
47. Moody, J., Saffell, M.: Learning to trade via direct reinforcement. IEEE Trans. Neural Netw. **12**(4), 875–889 (2001)
48. Lu, D.W.: Agent inspired trading using recurrent reinforcement learning and lstm neural networks. arXiv:1707.07338 (2017)
49. Jiang, Z., Xu, D., Liang, J.: A deep reinforcement learning framework for the financial portfolio management problem. arXiv:1706.10059 (2017)
50. Zhang, Y., Zhao, P., Wu, Q., Li, B., Huang, J., Tan, M.: Cost-sensitive portfolio selection via deep reinforcement learning. IEEE Trans. Knowl. Data Eng. **34**(1), 236–248 (2020)
51. Bäuerle, N., Rieder, U.: Markov Decision Processes with Applications to Finance. Springer, Berlin (2011)

52. Murphy, J.J.: Technical analysis of the financial markets: a comprehensive guide to trading methods and applications (1999)
53. Huang, C.Y.: Financial trading as a game: a deep reinforcement learning approach. arXiv:1807.02787 (2018)
54. Busseti, E., Ryu, E.K., Boyd, S.: Risk-constrained kelly gambling. arXiv:1603.06183 (2016)
55. Lee, N., Moon, J.: Offline reinforcement learning for automated stock trading. IEEE Access (2023)
56. Snoek, J., Larochelle, H., Adams, R.P.: Practical bayesian optimization of machine learning algorithms. Adv. Neural Inf. Process. Syst. **25** (2012)
57. Gu, J., Du, W., Rahman, A.M., Wang, G.: Margin trader: a reinforcement learning framework for portfolio management with margin and constraints. In: Proceedings of the Fourth ACM International Conference on AI in Finance, pp. 610–618 (2023)
58. Soleymani, F., Paquet, E.: Financial portfolio optimization with online deep reinforcement learning and restricted stacked autoencoder-deepbreath. Expert Syst. Appl. **156**, 113456 (2020)
59. Thakur, M., Kumar, D.: A hybrid financial trading support system using multi-category classifiers and random forest. Appl. Soft Comput. **67**, 337–349 (2018)
60. Li, Y., Du, N., Song, X., Yang, X., Cui, T., Xue, N., Farjudian, A., Ren, J., Cheah, W.P.: Cardinality and bounding constrained portfolio optimization using safe reinforcement learning. In: 2024 International Joint Conference on Neural Networks (IJCNN), pp. 1–8. IEEE (2024)
61. Sharpe, W.F.: The Sharpe ratio. Streetwise Best J. Portf. Manag. **3**, 169–185 (1998)
62. Young, T.W.: Calmar ratio: a smoother tool. Futures **20**(1), 40 (1991)
63. Sortino, F.A., Price, L.N.: Performance measurement in a downside risk framework. J. Invest. **3**(3), 59–64 (1994)
64. Keating, C., Shadwick, W.F.: An introduction to omega. AIMA Newslett. (2002)

Authors and Affiliations

Himanshu Choudhary[1] · Arishi Orra[1] · Kartik Sahoo[1,2] · Manoj Thakur[1]

✉ Kartik Sahoo
kartiksahoo@soa.ac.in

Himanshu Choudhary
d21024@students.iitmandi.ac.in

Arishi Orra
d21022@students.iitmandi.ac.in

Manoj Thakur
manoj@iitmandi.ac.in

[1] School of Mathematical and Statistical Sciences, Indian Institute of Technology Mandi, Mandi, Himachal Pradesh 175005, India

[2] Department of Computer Science and Engineering, Siksha 'O' Anusandhan (Deemed to be) University, Bhubaneswar, Odisha 751030, India