International Journal of Applied Mathematics

Volume 38 No. 10s, 2025

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

# A REVIEW OF ADVANCEMENTS AND APPLICATION OF THE STOCK MARKET IN Q-LEARNING REWARD FUNCTION ANALYSIS

**Suchita Nilesh Borkar*[1], Dr.Sheetal Bansude Bura[2], Vaishali Vitthal Hadawale[3], Dr. Jayashree Vispute[4]**

[1]Assistant Professor

Pimpri Chinchwad College of Engineering,Pune

suchitaborkar1@gmail.com

[2]Assistant Professor,

School of Commerce and Management D.Y.Patil International University,Akurdi,Pune.

Sheetal.bura@gmail.com

[3]Assistant Professor

D.Y Patil College of Agriculture Business Management,Akurdi,Pune.

vaishali_bharat@rediffmail.com

[4]Associate Professor

Faculty of Commerce and Management Vishwakarma University ,Pune.

jayashree.vispute@vupune.ac.in

**Abstract**

Deep Q-learning is a common reinforcement learning approach in this technique including training of this method using deep neural networks (DNN), in this technique also using Deep Q-networks (DQN) network approximates widely recognized Q-functions. The popular version of deep Q-learning is developed under practical and verifiable principles by a dynamical viewpoint, which also offers a theoretical analysis, despite the lack of formal guarantees that would limit its usage in reality. Algorithms are analyses of convergence of the asymmetric behavior of the educational process. Stock exchange forecasting is now in high demand, and most investors predict challenging tasks, researcher's analysts of the financial market are noisy, volatile, nonparametric, complex, nonlinear. Technology used Deep reinforcement learning (DRL) algorithms previously to find intractable problems. DRL makes it possible to generate automation profit in the stock market financial combining price assets prediction steps and allocation steps used in this portfolio method is a unified process is a fully autonomous system capable of environment interaction to make the decisions optimal through fault and trial. The paper focuses on showing various reward functions used for learning patterns and making optimum decisions.

**Keywords: Deep Reinforcement Learning, Q-learning, Deep Learning Model, Deep Q-Network Model.**

## 1 INTRODUCTION

Machine learning is used to explore, grow in nature, become an oasis, and investigate the most difficult applications using trading stock. This technique is recognized as a decision-making procedure that facilitates the optimization of value in trading, allowing for the return of excess investments to enhance

Received: August 10, 2025

Volume 38 No. 10s, 2025

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

metrics related to profitable economic services, while also adjusting for risk and return, among other factors [1]. Stock market trading is growing to popularize and become a rich field exploring powerful computing technology is fast to develop exploration tools [5]. Widely used in the system, the trading strategies are better providing performance problem metrics to maximize the profit value, among explored, economic utility [6]. A machine learning solution for stock trading emerged even before the widespread adoption of computers among investors, thereby diminishing the reliance on human assistance [7]. Machine learning results are based upon Technical Analysis (TA) has been widely used to propose the financial forecasting literature it is used in deep learning support vector machine random forest, and reinforcement learning, among others [10]. Study tries to build a prediction that is an effective model in the stock market [3]. Reinforcement learning interacts with environments to learn the policy is optimal through fault and trials are found in the decision-making problem in various frameworks

## 2 Reinforcement learning

In Reinforcement learning (RL), algorithms are used in learning intelligence agents through the problem and trial, in designing the environment, and in this reward to get the output [25]. The agents give the task of maximizing the discounted environment to give the reward. Policy iteration and Value are used in an agent's intelligence [26] Q-learning types are used in the critic method. DQN is an exploit replay algorithm that successfully integrates RL in a deep neural network. The current investigation progress of the DRL model shows the possibility of developing the stock price accuracy of the trend forecast using the DRL and designing the deep neural network [29]. **1**Reinforcement learning between agent and the environment This method, stock market trends costs are predicted for the transaction [30].

Reinforcement Learning (RL) sequential decision makes the agent interact with an environment to optimize the cumulative reward signal [37]. Interact occurs on a trial-error basis, with the agent learning good behavior in past experiments. In RL is to describe discrete time stochastic control processes [32].

**Table 1: Compare sample-based RL and sample-free RL**

| **Elements** | **sample-based RL** | **sample-free RL** |
|---|---|---|
| The quantity of iteration among environment and agent | Smaller | Bigger |
| Convergence | Quick | Delay |
| Prior knowledge of transition | agreed | never |
| flexibility | Strong depending upon the learning model. | Adjust by error and trial. |

The trading algorithm is a prime interest of many companies' sole technique to make trading safe. The stock market companies for the enthusiast and also extensively in research area on machines is to predict a stock market [8].

## 3 Q-Learning

The two essential forms of the Q-learning version are synchronous and asynchronous counterparts [38]. In synchronous Q-learning, it is generally assumed to access the simulator and generate an independent sample of all states and action pairs, with an attempt to update all other entries in the Q-function evaluated simultaneously [1]. In the study of asynchronous structure in Q-learning, it is naturally used on

Received: August 10, 2025

**International Journal of Applied Mathematics**

**Volume 38 No. 10s, 2025**

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

data in Markovian trajectory to include the behavior policy. Although the most prior and sufficient frequency [2]. In recent work, this sample complex on asynchronous structure in Q-learning is most of the subsequence or shared by the worth on some variants on a model-free algorithm, for example, a variant couple in upper confidence bounds has been proven the effective result in online exploratory [3].

**Q-function values**

The function value is to apply and evaluate the agent to utilize the policy $\pi$ to visit $t$, with "good" determining the term of expected return. The result is expected to be received in the future and it depends upon the action taken [3]. Mathematically it uses the function as the expected return the function is approximately through the Bellman expectation $X^{\pi}(t_u)$ is the expected return and $t_u$ is stated at the time $u$, if $\pi$ under the policy is expressed equation (1)

$$X^{\pi}(t_u) = \mathrm{E}\left[s_{u+1}\gamma X^{\pi}(t_{u+1})\right] \tag{1}$$

Where $X^{\pi}(t_u, b_u)$ is the expected return of taking action $b_u$ and the state $t_u$, if the statement $t_{u+1}$ the state at the next time step $u+1$, $E$ is the environment from which the next state $t_{u+1}$ and reward received after transitioning from the state $s_{u+1}$ $\gamma$ It is the discount factor. And the return statement is implemented in equation (2).

$$X^{\pi}(t_u) = \sum_{bu\in \mathrm{A}} \pi(b_u \mid t_u) \sum_{tu+1\in t} U(t_{u+1} \mid t_u, b_u)\left[S(t_u, t_{u+1}) + \gamma X^{\pi}(t_{u+1})\right] \tag{2}$$

Where $\sum_{bu\in \mathrm{A}}$ it represents a summation of all possible action $b_u$ s it belongs to action space $\sum_{tu+1\in t} z$ and represents all possible next states $t_{u+1}$ in which the agent could transition to form the current states $t_u$ , $u$ at the time, $U(t_{u+1} \mid t_u, b_u)$ it transitioning probability represents the from one state $t_{u+1}$ to the next state $t_{u+1}$ after taking action $b_u$ and $S(t_u, t_{u+1})$ it gives instant reward is receive for transitioning from $t_u$ state to $t_{u+1}$ states.

This equation is commonly referred to as the Bellman equalization. The agents always select the action according to the policy $\pi^*$ it maximizes to evaluate the Bellman equation is expressed in equations (3) and (4)

$$X^*(t_u) = \max_{bu} \sum_{tu+1\in t} U(t_{u+1} \mid t_u, b_u)\left[S(tu, t_{u+1}) + \gamma W^*(t_{u+1})\right] \tag{3}$$

$$\stackrel{\Delta}{=} \max_{bu} Q^*(t_u, b_u) \tag{4}$$

Where $\max_{bu}$ the action $b$ maximizes the Q-value, although the equation is obtained by an optimal value function $X^*$ it does not provide the information is not enough to re-construct the optimal policy $\pi$. Thus

Received: August 10, 2025

**International Journal of Applied Mathematics**

**Volume 38 No. 10s, 2025**

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

the function is used to evaluate how good an agent is at performing a particular action $(b_u)$ thus the state $(t_u)$ it the policy $\pi$ with also implemented in equation (5)

$$Q^{\pi}(t_u, b_u) = \sum U(t_{u+1} \mid t_u, b_u)\left[S(t_u, t_{u+1}) + \gamma X^{\pi}(t_{u+1})\right] \tag{5}$$

The unlike value specifies the goodness state on the Q-function; it specifies the action on its state.

## 5 Reward function

*Reward shaping* uses domain knowledge to create a more complex reward function. Reward shaping aims to accelerate learning and overcome problems in exploration and credit assignment when the environment provides insufficient or delayed rewards [50]. The reward shaping framework entails substituting the actual function with the function that is dynamically updated as the reinforcement learning (RL) agent interacts with the framework. This approach does not necessitate human guidance .

The potential-based framework indicates that variations in the rewards do not influence the optimal policy [6]. Potential-based shaping involves potential functions based on state and action pairings, rather than using only states. The RL algorithm does not rely on and uses the distance-to-goal reward and learns the dense reward used asymptotically leading to optimal policy in the original sparse reward [6].

*A sparse* reward environment is used in various ways [7]. In an environment characterized by sparse rewards, the task is divided into smaller subtasks, allowing the system to develop higher-level strategies through auxiliary policies that enhance exploration [8]. The algorithm designed to learn and refine these methods, alongside the original reward structure, varies from existing approaches.

In reward calculation analysis various stock market applications in reinforcement learning. Stock market applications use many reward functions, reward calculation formulas, advantages, and challenges. Profit-based reward, Profit and Loss and Sharpe ratio reward, Proposed reward, Reward shaping using price trailing, Conditional Sharpe ratio, Risk adjustment return, Transaction costs, Cumulative return, Fitness function, PSR reward, NAV reward, Profit rate.

Table 2 all reward calculation

| **Reference** | **Reward function** | **Reward calculation formula** | **Advantages** | **Challenges** |
|---|---|---|---|---|

Received: August 10, 2025

**International Journal of Applied Mathematics**

**Volume 38 No. 10s, 2025**

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

| | | | | |
|---|---|---|---|---|
| [11] | Profit based reward | The profit based reward is defined as $S_u^{(profit\ and\ loss)} = \begin{cases} z_u, & agent\ is\ long \\ -z_u, & agent\ is\ short \\ 0, & agent\ has\ no\ positior \end{cases}$ Where $S_u^{(PnL)}$ is the reward at a time $u$ for the agent based on profit and loss, $z_u$ is positive, the agent made a profit, if $-z_u$ is negative the agent incurred a loss, if the agent does not position the reward $is\ 0\ is\ 0$ indicating no profit or loss. | The reward agent is based on profit and the position is taken in a common methodology for trading in reinforcemen t learning. | The reward profit and loss take time to appear making it hard for the agent to specific actions in their outputs. |
| [12] | PnL and Sharpe ratio | Overall profit and losses are defined and Sharpe ratio reward is also defined. $\begin{cases} S_u^{(PnL)} + S_u^{(fee)}, & if\ u < x \\ S_u^{(PnL)} + S_u^{(fee)} + S_u^{(Sharpe)}, & if\ u \geq x \end{cases}$ Where $u$ is typically represents time, $u < x$ only compounds are considered (PnL and fee) and if $u \geq x$ the Sharpe ratio reward is added, if the $S_u^{(PnL)}$ is the profit and loss at time $u$ it reflects the returns from trading if $S_u^{(fee)}$ it represents the fee incurred at the time $u$ and includes transaction costs, management fees, in the $S_u^{(Sharpe)}$ term represents the Sharpe ratio reward at time $uu$, Sharpe ratio is calculated as $Sharperatio = \frac{F[S] - S_g}{\sigma}$ Where $F[S]$ is the expected return from the portfolio, if $S_g$ is risk free rate, typically the return on government bonds and $\sigma$ represents the standard deviation of returns representing risks. | In RL episodes are calculated over a short period. | In this reward is a metric calculated annually |

Received: August 10, 2025

**International Journal of Applied Mathematics**

**Volume 38 No. 10s, 2025**

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

| | | | | |
|---|---|---|---|---|
| [11] | Reward shaping using price trailing | Where $n$ the margin fraction parameter is used to calculate the distance between the margin and the most recent target price and the agent's purpose is to keep the agent price as $q_b(u)$ near the target price as possible. The reward trailing is established. $s_u^{(trail)} = 1 - \frac{\lvert q_b(u) - q_d(u) \rvert}{nq_d(u)}$ Where $s_u^{(trail)}$ it represents the trailing reward at time $u\,u$ it measures how close the agent price is to the target price, if $q_b(u)$ it represents the agent price at time $u\,u$ it is the price that the agent is trying to adjust to match the target price. I $q_d(u)$ t is the target price at the time $u$. The agent aims to keep its price close to this target and $n\,n$ is the margin factor parameter. It determines the acceptable distance between the agent price and the target price. | It reduces unnecessary transactions. | The reward function is to adapt to the sudden market shift |
| [13] | Conditional Sharpe ratio | The conditional value at risk $CVaS$ reward ratio is defined as $Conditional\ sharperatio = \frac{(s - s_g)}{CVaS_{1-\alpha}}$ Where $s$ is a fund returns, $s_g$ is a risk free rate, $CVaS_{1-\alpha}$ is a value conditional risk over the specified time horizon $100(1-\alpha)\%$ at the confidential level. | The conditional Sharpe ratio is the ratio to expected excess return over the risk-free rate. | - |

Received: August 10, 2025

**International Journal of Applied Mathematics**

**Volume 38 No. 10s, 2025**

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

| | | | | |
|---|---|---|---|---|
| [14] | Risk adjustment return | The RAR of the hold and buy methods $RAR_{H\&B}$ is calculated conditional Sharpe ratio and is defined as $RAR_{B\&H} = \frac{R_{H\&B} - s_g}{CVaS_{1-\alpha}}$ Where $RAR_{H\&B}$ adjust return for hold and buy, if $RAR_{H\&B}$s the return of the hold and buy strategy, $s_g$ its risk free rate in the excess return is calculated by subtracting $s_g$ from the return of the hold and buy, if $CVaS_{1-\alpha}$ is conditional value at risk at the confidential level $1-\alpha$. | The excess return of the trading rule over the buy-and-hold strategy of the fitness function | The risk-adjusted return used for buy and hold strategy |
| | Transaction costs | Where $R_{H\&B}$ is expected is the return of the hold and buy approaches. It calculates the dividends, splits, and transaction fees. $CVaS_{1-\alpha}$ is calculated $CVaS_{(1-\alpha)} = -F[Y \mid Y \le -VaS_{(i-\alpha)}]$ Where $F$ is the expected value, $Y$ is the stock's profit or loss over the provided risk horizon, and $VaS_{(1-\alpha)}$ is the lowest projected loss over the given time horizon at $1-\alpha$ confidential levels. $VaS_{(1-\alpha)}$ It approximates the historical simulation. | Investors intending to optimize the net return require compensation for investing in their securities with high transaction costs. | The common consensus is that greater transaction costs have high projected profits but lower trading volume. |
| | Cumulative return | The cumulative risks adjusted returns of trading methods are some of the RAR of all transactions in the period calculated. $RAR_{S\&B} = \sum_{j=1}^{o} \frac{R_j - s_g}{CVaS_{1-\alpha}}$ Where $RAR_{S\&B}$ the risks adjusted returns generate the trading rules. If the number of transaction periods advised by the trading rule is represented $o$, if $R_j$ transaction returns are considering the dividends, splits, and price transaction, and if $s_g$ is risk free rates and refers to the return of an investment that is considered free from risk for example government bonds. | The cumulative return used to the trading rules and the sum of the transaction return is considered dividends. | - |

Received: August 10, 2025

**International Journal of Applied Mathematics**

**Volume 38 No. 10s, 2025**

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

| | | | | |
|---|---|---|---|---|
| | Fitness function | The F function is calculated as:<br>$fitnessfunction = RAR_{S\&B} = RAR_{H\&B}$<br>The two constraints apply to trade rule evaluation: the first one is the sell signals are to implement the stocks that exist in the trader portfolio market, and the second one is the buy signal is to implement sufficient capital is available. | The corrections include fixing the bug in the script and estimating the probabilities used in the calculation of the model. | - |
| [15] | PSR reward | Agent reward with numerical points is based upon the quality of the action on each step. The proposed reward is capable of information by the agent, uses short-term daily returns and it is made after buying, selling, and holding decisions over multiple shares. The proposed reward function is defined below.<br>$\text{Reward} = change\ in\ portfolio + Sharperatio + 0.9 * daily\_returns$<br>The Sharpe ratio is given below.<br>$Sharperatio = \frac{F[S_b - S_c]}{\sigma_b}$<br>Where $F$ is the expected value, $S_b$ is asset return, $S_c$ is risk free returns, and $\sigma$ indicates a standard deviation of asset excess returns. And value 0.9 is tuned to carry out different simulations. | The agent is rewarded with numerical points based on the quality of their actions at each step. | PSR provides a comprehensive evaluation of portfolio outcomes, managing risk and investment plan efficacy |

Received: August 10, 2025

**International Journal of Applied Mathematics**

**Volume 38 No. 10s, 2025**

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

| | | | | |
|---|---|---|---|---|
| [16] | Net Asset Value( NAV) reward | In the environment, the reward is implemented for each stage of the training methods, and for each step is a market trend reversal noticed throughout time. The Net Asset Value (NAV) reward consists of two different compounds: NAV change reward and current NAVR. NAV change reward is expressed in<br>$\text{Reward} = \frac{\text{changed composition} - \text{orginal composition}}{\text{orginal composition}} \times TSF$<br>Where a change in the composition refers to the NAV acquired after 10 days based on portfolio composition changes caused by RL activities, the agent is the initial portfolio composition and TSF indicates the time scaling factor. In the technique is difference between the original composition and the changed composition is visualized at the time $u = 2$ and the term of value is determined by TSF.<br>$TSF = 0.5 \times \frac{\text{Number of days past}}{\text{Total number of days}} + 0.5$<br>In the TSF above the equation adjusts the NAV change reward from 0.5 to 1.0 lowering the degree of rewards available with time. This is to emphasize the importance of initial activities, which have a higher impact on the end NAV through compounding. The second component, the current NAV reward, is computed by simply dividing the current NAV by a constant of 10000000 to normalize the given reward.<br>$\text{Current NAV reward} = \frac{\text{Current NAV}}{1 \times 10^7}$<br>Where the current NAV is the NAV obtained after 10 days with the modified portfolio composition. It is adjusted to match the current NAV. The total reward for each step is calculated as<br>$\text{Total reward} = \text{NAV change reward} + \text{current NAV reward}$<br>In reinforcement learning, the agent gains knowledge of the performance of immediate action depending on the present condition of the environment. | In this technique, RL agents show a better percentage of Net Asset Value max drop and much lower volatility. | - |

Received: August 10, 2025

**International Journal of Applied Mathematics**

**Volume 38 No. 10s, 2025**

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

| | | | | |
|---|---|---|---|---|
| [17] | Profit's rates | The selling agents have a few additional bits in their state to indicate the profit rates achieved from the holding stocks during an episode. The profit rates on Day E, $$PR_E = \frac{P_E^C - BP}{BP} \times 100$$ Where value encodes the $PR_E$ fixed length, $C$ current price, and $BP$ is the buy price. | The single episode ends after the framework notifies the buying stock signal agent of the result profit rates at the rewards. | The selling signal agent has a few more bits in the state to represent the profit rates obtained by holding the stock price during an episode. |
| [18] | DDPG reward | Assume $o, t_u \in S^o \; Q_u \in S^o$, $Bwd_u$ it indicates the sum of trading stocks, quantity of stocks owned in the networks, available cash at the time $u$ and overall portfolio value $W_u$ is equal in $$W_u = Bwd_u + t_u \times Q_u$$ The $PL_u$ at-time $u$ and next-time step $u+1$ $$PL_u = W_u - W_{u-1}$$ The calculated in different portfolio values at time risk free returns is measured by return is defined as $$RS_u = \frac{\sum_{u \in 1}^{U} QM_u - GS_u}{\delta QM_u}$$ Where $RS_u$ is the Sharpe ratio, at the time $u$, it $QM_u$ is the Q-value at the time $u$ and age return earns more than the risk-free return $GS_u$ for the same period. The volatility is represented as the average deviation of the P and L time interval $\delta$ s. | The policy-based model generates continued action on each interval. An agent can purchase or sell any component of an asset under the same rules as a discrete action. | In available funds to perform the buying process, it is not executed. |

The calculation is used for actions that lead to profit or loss and is expressed in equation (6); if an action leads to profits [21]. The value of this technique corresponds to the profits generated; conversely, if an action results in a loss, the value of the rewards is three times the amount of that loss. This encompasses the total of all rewards obtained throughout the episodes.

Received: August 10, 2025

**International Journal of Applied Mathematics**

**Volume 38 No. 10s, 2025**

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

$$\text{Reward} = \begin{cases} profit, & in\ case\ of\ profit \\ 3\times loss, & in\ case\ of\ loss \end{cases} \quad (6)$$

Where $profit$ if an action leads to profit and $3\times loss$ the reward value of the profit is equal

In algorithmic trading scenarios, employing reinforcement learning (RL) reward strategies based on daily returns is a logical approach.

## 6 DISCUSSIONS

Reinforcement learning (RL) is used to apply the stock market Q-learning depending upon the Q-function to estimate the future rewards it gives the state action pair. The rewards play a crucial role in training agents to optimize decisions such as buy, sell, or hold stock, and various reward calculation types that are dense, sparse, shaping, penalty, and hybrid are used in the agent learning process.

*Adaptive sentiment–aware DDPG approach* the evaluation of the learning agent and the portfolio values occurs daily [34], with results obtained from the baseline method of adaptive DDPG, alongside traditional mean and mean-variance analysis. This approach quantifies the value of the stocks held by the trader and assesses the model's performance in comparison to standard benchmarks. By integrating the stock's price owned, a remaining neutral of a trading agent, there was a significant enhancement in the trading agent's performance.

*Dense reward* systems enable quick learning by providing instant feedback; it is also used to lead to overfitting and excessive trading . The agent becomes too sensitive to short-term price changes to the frequency caused by the noise rather than real patterns. This is particularly troublesome in the stock market, where transaction costs may quickly destroy earnings from frequent trading.

*Sparse reward structures* emphasize long-term profitability and are used to allow the agent to focus on large deals [26]. This method is slow learning due to infrequent feedback, complicating the credit assignment problem. Agents may fail to connect the previous actions to delayed effects making it difficult to discover optimum methods in unpredictable contexts.

*Shaping reward* Researchers can assist agents throughout the early stages of learning by delivering intermediate feedback rewards that do not overwhelm them. In the agent, care must be taken to ensure that the shaped rewards do not result in poor behavior [27]. The design of the shaped rewards should be aligned with the ultimate goals of the trading strategy to avoid diverging in intended outputs.

*Hybrid reward* systems take advantage of both dense and sparse structures in the agent allowing for more adaptable learning [28]. By offering small rewards for achieving milestones while maintaining the focus on long-term profitability, these systems can motivate agents to try other techniques. Future studies should look at ideal hybrid system designs to improve both learning efficiency and profitability [29].

*Risk management* the tendency of agents to overtrade in dense reward environments emphasizes the need for measures that encourage sensible decision-making [89]. Penalty rewards can successfully prohibit poor trading behaviors, and manage to prevent undue caution from impeding profitable deals [31].

Table 4 Comparison of reward function calculation Q-function and Reinforcement learning in the stock market

| | **Methods** | **Applications** | **Dataset type** | **Analysis** | **Merits** | **Demerits** | **Reward calculation** |
|---|---|---|---|---|---|---|---|

Received: August 10, 2025

**International Journal of Applied Mathematics**

**Volume 38 No. 10s, 2025**

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| [21] | Deep Q-Network Trading | Algorithmic trading to determine optimal trading (long or short ) in stock market | Historical stock market data | Performance assets focused on the Sharpe ratio | DRL uses promising performance and rigorous assessment methodology | Limited by the quality and quantity of historical data, potential overfitting | The rewards are based on the daily returns of the trading strategy to maximize the predicted discounted total of the payment over an unlimited time horizon |
| [32] | Nested RL, Weighted Random Selection with Confidence (WRSC) | Stock trading, investors in portfolio management | Historical stock data | The method is evaluated by using performance indicators like annualized return, cumulative return, Sharpe ratio, and MDD | High returns, better risk management, dynamic adoption to market change | Complexity in implementation, potential overfitting | This framework is based upon changing the different portfolio values, the log value of the portfolio change ratio |
| [33] | Multi-agent Q-learning | The framework is applied to the KOSPI 200 stock market | Historical stock prices | The performance is evaluated based on profitability and risk management, and the results are compared to a supervised learning-based system. | High profits, better risk management, and effective integration of long-term and short-term trading strategy | Complexity in implementation and need for extensive training episodes | Rewards are based on the profit rate for buy signals and the success rate of trades for calculating rewards based on stock price movements |

Received: August 10, 2025

**Volume 38 No. 10s, 2025**

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| [12] | Sharpe ratio-based reward shaping approach, DRL, PnL | Financial trading | Cryptocurrency trading date | The proposed method is evaluated through experiments, showing the improved risk-adjusted performance of the trading agents compared to using PnL or Sharpe ratio-based rewards. | The agent's ability to manage risk leads to better overall portfolio performance. | The method's effectiveness depends upon the accuracy of the Sharpe ratio | The reward combines PnL and Sharpe ratio, the adjustment based on the agent performance in each step, aiming to achieve a higher Sharpe ratio |
| [11] | Deep reinforcement learning, price trailing-based reward shaping | FOREX trading aims to develop a market-wide RL agent that can trade multiple currency pairs effectively | FOREX trading | The approach is evaluated using this metric Profit & Loss, Sharpe ratio and maximum drawdown show significant improvements in trading performance. | The method improves profitability and stability. | It requires careful design of rewards and preprocessing to avoid overfitting. | Rewards are shaped using the price trailing method, in reward distribution reduces the noise from PnL-based rewards. |
| [15] | A multifaceted approach to the stock market using RL | Multi-stock environment using algorithm Advantage Actor-critic(A2C) and Deep Deterministic policy gradient (DDPG) | Daily stock data, from Dow Jones 30 companies. | The framework is validated through backtesting and shows better performance compared. | The model provides a comprehensive view of the market, improving decision-making. | High volatile markets and require extensive data for training | A Portfolio Sharpe ratio reward function is proposed, considering portfolio value, Sharpe ratio, and daily returns to guide the RL agent decision |

Received: August 10, 2025

Volume 38 No. 10s, 2025

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| [24] | A deep reinforcement learning approach that incorporates market sentiment for social media | Industrial aims to maximize returns while minimizing risk. | Market sentiment and Historical stock prices | The method compares against the traditional approach using metrics like the Sharpe ratio and annualized investment return. | The approach is more robust and adaptive | Extensive data processing and computational resources | The reward is based on changes in portfolio value and market sentiment, with positive rewards for profitable actions and sentiment–aligned decisions |
| [23] | Deep Q learning for stock trading | Trading technique | Daily stock prices | The performance of different models is evaluated based on profitability and volatility. Deep Q-learning showed higher average profits but also. | Generates positive profit strategies are directly trading stocks without further optimization | High volatility requires careful tuning and large datasets | Rewards are used in this method to give profit for each action (buy, hold, sell) |

## 7 PERFORMANCE ASSESSMENT

In stock market applications Q-function and reinforcement learning calculate the reward function used in the model and evaluation metrics are used for calculation. The stock market uses Q-function and reinforcement learning in many paper stock trading are used, finance used in the reward calculation and applications are used in Q-function models are used in applications used models some B&H and S&H are used in stock trading, in this technique are used trading algorithms are TF, MR, TDQN, and this model are used in portfolio investor trading, A2C, DDPG, DDPG RL, WRSC, Baseline^ DJI index, and Sentiment. Reward, Buy and Hold, DQN-agent and evaluation reward functions used in this calculation are Sharpe ratio, Annualized return, Max drawdown, Sortino ratio, and Calmar ratio.

Table 5 Overall performance indicator

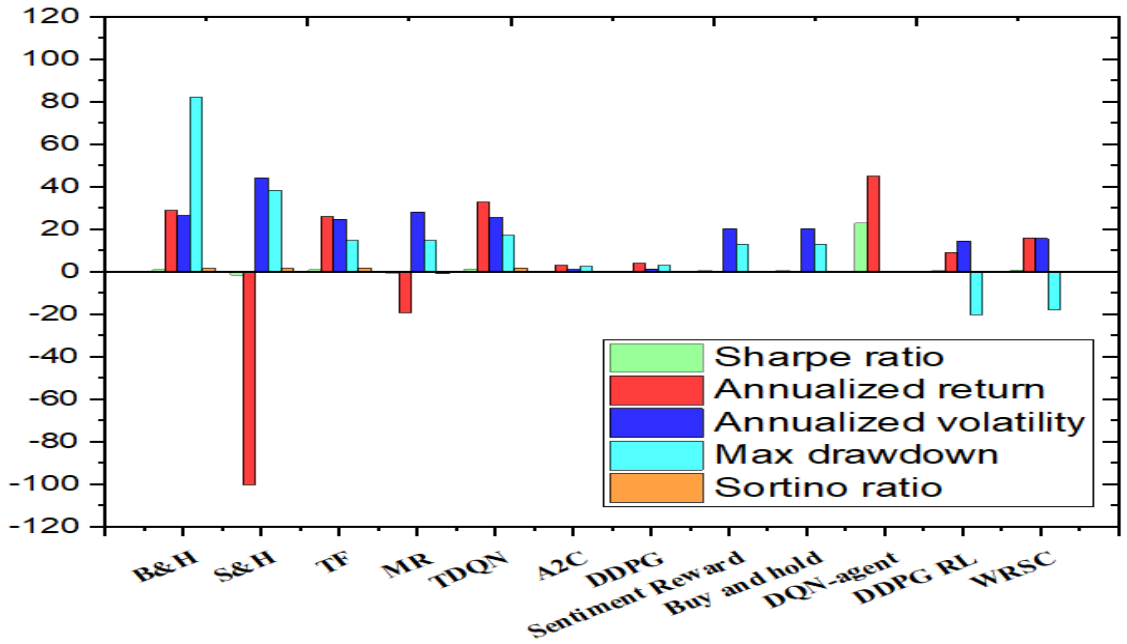
| **reference** | **model** | **Sharpe ratio** | **Annualized return** | **Annualized volatility** | **Max drawdown** | **Sortino ratio** |
|---|---|---|---|---|---|---|
| [21] | B&H | 1.239 | 28.86 | 26.62 | 82.48 | 1.558 |

Received: August 10, 2025

**International Journal of Applied Mathematics**

**Volume 38 No. 10s, 2025**

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)




| | | | | | | |
|---|---|---|---|---|---|---|
| | S&H | -1.593 | -100.00 | 44.39 | 38.51 | 1.558 |
| | TF | 1.178 | 25.97 | 24.86 | 14.89 | 1.802 |
| | MR | -0.609 | -19.09 | 28.33 | 14.89 | -0.812 |
| | TDQN | 1.484 | 32.81 | 25.69 | 17.31 | 1.841 |
| [15] | A2C | 0.15 | 3.04 | 1.122 | 2.479 | 0.22 |
| | DDPG | 0.23 | 4.03 | 1.44 | 3.03 | 0.08 |
| [34] | Sentiment Reward | 0.76 | - | 20.43 | 13.02 | - |
| [35] | Buy and hold | 0.72 | - | 20.43 | 13.02 | - |
| [8] | DQN-agent | 23.07 | 45.05 | - | 4.35% | - |
| [36] | DDPG RL | 0.68 | 9.23 | 14.46 | -20.04 | - |
| | WRSC | 1.02 | 16.01 | 15.8 | -17.76 | - |

## PREDICTION MODEL

The bar chart provided the comparison of the models across different metrics, Mean Reversion (MR) and sentiment B&H perform well across multiple metrics, with MR excelling in risk-adjusted return and sentiment B&H in annualized returns. Stability Buy and Hold show low annualized volatility and maximum drawdown indicating more stable and less risky returns. And worst performance WRSC model generally performs the worst across most metrics with the lowest Sharpe ratio, highest drawdown, and lowest annualized return,

Received: August 10, 2025

**International Journal of Applied Mathematics**

**Volume 38 No. 10s, 2025**

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

**8 EVALUATION METRICS**

*Sortino ratio* (SR) is considered the better metric that is accessed by showing the trade and also offers a look at the expected profit when risk is applied. The equation is expressed in (16)

$$Sortino\ ratio = \frac{\bar{s} - s_k}{\sigma_e} \tag{16}$$

Where expected average returns of the strategy $\bar{s}$, if the risk free rates finally $s_k$ and $\sigma_e$ the downside deviation of the return is unexpected.

*Maximum Drawdown* (MDD) is the calculation to quantify the price of risks through the period under consideration. If lower values are maximized the drawdown is usually preferred. It $Q$ is a find the values before a large dropping and $W$ is the price before the new high price is incorporated in equation (17)

$$MDD = \frac{(Q - W)}{Q} \tag{17}$$

*Annualized return*s assess the effectiveness of the trading stock technique, utilizing the annualized return measurement index, as investors prioritize overall investment advantages regardless of the strategy used. The annualized return is the rate earned by investors, in a one-year investment period are calculated formula is expressed in equation (18)

$$Annualized return = \frac{return.365}{principal.investment days} \tag{18}$$

*The cumulative* return reflects the total amount of investment that has either been lost or gained over a period, irrespective of the duration involved this return may be computed using the formula given in equation (19) and should be reported as the percentage.

$$cummulative return = \frac{current values - orginal values}{orginal values} \tag{19}$$

*Sharpe* ratio evaluates the performance metrics of investments by comparing it to a risk free asset, taking into account the risk associated. It is computed by taking the difference between the returns on the investment and the return that is free from risk; it is subsequently divided by the standard deviation of the investment. The formula for calculating the Sharpe ratio is presented in equation (20).

$$T_b = \frac{F[S_b - S_b]}{\sigma_b} = \frac{F(S_b - S_b)}{\sqrt{\text{var}[S_b - S_b]}} \tag{20}$$

Where $S_b$ the asset is returned and $S_b$ is the risk-free return. $F[S_b - S_b]$ Is an excess asset return over the benchmark likely to have a value, $\sigma_b$ its standard deviation

*Accumulated Wealth Rate* (AWR) is the total profit and loss expressed in equation (21). Throughout the trial trading phase U, is split by the total amount of cash on hand at the beginning of the trade $Bwd_{u=0}$.

Received: August 10, 2025

**International Journal of Applied Mathematics**

**Volume 38 No. 10s, 2025**

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

$$AWR_U = \frac{\sum_{u\in 1}^{U} PL_u}{Bwd_{u=0}} \tag{21}$$

*Average Profit Return (APR) i*s characterized by the mean of the yearly return rates, which is determined by calculating the average AWR achieved for every year during the specified test time, as represented in equation (22).

$$APR_U = \frac{\sum_{u\in 1}^{U} AWR_u}{U} \tag{22}$$

*Average Sharpe ratio* (ASR) the average of the ASR during the test trading time is known as the average SR, or ASR equation (23) expresses the SR metric

$$ASR_U = \frac{\sum_{u\in 1}^{U} (SR_u)}{U} \tag{23}$$

*The Average Maximum Draw Down (AMDD)* refers to the largest decrease in the value of a trading portfolio, measured during a given time measure U, from its highest point to its lowest position. The portfolio value at any given moment is denoted as $w_u$ , and the AMDD is calculated as the mean of the annual Maximum Draw Down (MDD) throughout the designated trading period, as represented in equations (24) and (25).

$$MDD_u = \frac{(\max w_u - \min w_u) u \in U}{w_u} \tag{28}$$

$$AMDD_u = \frac{\sum_{u}^{U} (MDD_u)}{U} \tag{29}$$

*Average Calmar Ratio* (ACR) the average profit return over a period divided by the MDD for that period is known as ACR, and equations (26) and (27) establish the ACR average of the annual (CR) during a test trading period $U$

$$CR_u = \frac{APR_u}{MDD_u} \tag{26}$$

$$ACR_u = \frac{\sum_{u\in 1}^{U} (CR_u)}{U} \tag{27}$$

## 9 FUTURE SCOPES

The stock market application in the future of Q-function reward calculation involves some reward functions that are dense, sparse, shaping, penalty, and hybrid rewards hold significant promise. Dense rewards can provide frequent feedback for small markets and allow more responsive trading. In sparse reward is better suited for long-term investment. Reward shaping is helpful for Q-function to guide learning towards a profitable market. The penalty reward is an effective managed risk. The hybrid reward is to combine both rewards and penalties to the balance approaches. The Q-function optimizes and

Received: August 10, 2025

**Volume 38 No. 10s, 2025**

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

analyzes more complicated real-time data and risk management in the Q-function is lowered, exploring market volatility RL is a rapidly growing area of deep learning research, while the cumulative rewards concern a single state action space, it would be interesting to investigate a situation with several agents.

## CONCLUSION

This review compares some reward calculations in the stock market and uses Q-learning and reinforcement learning in stock market applications to solve some reward calculations. Some reward functions used in the stock market domain are dense, sparse, shaping, penalty and hybrid rewards. Profit & loss are commonly used reward functions and portfolio reward functions in the stock market and many reward functions are used in finance, trading, and the stock market. Machine learning techniques cover the crucial role of many finance contexts, assess possible situations, and the technique for assessing investment risks to minimize losses brought on by fault strategies. To find the best trading position at any given moment during stock market trading activity, the deep Q-N algorithm is used in DRL to solve the algorithm trading problem. TDQN algorithm outperforms the standard technique in several ways. The cash basis reward function outperforms the accrual basis reward function in double DQN training with higher returns. In stock market trading integration strategies are used for reinforcement learning. Our solution uses layered RL to produce multiple DRL agents, such as A2C and DDPG approach the strategy in three RL agents to compute the maximum.

## REFERENCES

1. Wu, Mu-En, et al. "Portfolio management system in equity market neutral using reinforcement learning." *Applied Intelligence* 51.11 (2021): 8119-8131.
2. Chandola, Deeksha, et al. "Forecasting directional movement of stock prices using deep learning." *Annals of Data Science* 10.5 (2023): 1361-1378.
3. Khattak, Bilal Hassan Ahmed, et al. "A systematic survey of AI models in financial market forecasting for profitability analysis." IEEE Access (2023).
4. Zhao, Yu, et al. "Stock movement prediction based on bi-typed hybrid-relational market knowledge graph via dual attention networks." *IEEE Transactions on Knowledge and Data Engineering* 35.8 (2022): 8559-8571.
5. Thakkar, Ankit, and Kinjal Chaudhari. "A comprehensive survey on portfolio optimization, stock price and trend prediction using particle swarm optimization." Archives of Computational Methods in Engineering 28.4 (2021): 2133-2164.
6. Rahmani, Amir Masoud, et al. "Applications of artificial intelligence in the economy, including applications in stock trading, market analysis, and risk management." IEEE Access (2023).
7. Zhang, Kaiqing, Zhuoran Yang, and Tamer Başar. "Multi-agent reinforcement learning: A selective overview of theories and algorithms." Handbook of reinforcement learning and control (2021): 321-384.
8. Fang, Zheng, Biao Zhao, and Guizhong Liu. "Image augmentation based momentum memory intrinsic reward for sparse reward visual scenes." *IEEE Transactions on Games* (2023).
9. Yan, Jiangyue, Biao Luo, and Xiaodong Xu. "Hierarchical reinforcement learning for handling sparse rewards in multi-goal navigation." *Artificial Intelligence Review* 57.6 (2024): 1-23.
10. Ramírez, Jorge, Wen Yu, and Adolfo Perrusquía. "Model-free reinforcement learning from expert demonstrations: a survey." *Artificial Intelligence Review* 55.4 (2022): 3213-3241.
11. Tsantekidis, Avraam, et al. "Price trailing for financial trading using deep reinforcement learning." IEEE Transactions on neural networks and learning systems 32.7 (2020): 2837-2846

Received: August 10, 2025

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

12. Rodinos, Georgios, et al. "A Sharpe Ratio based reward scheme in Deep Reinforcement Learning for financial trading." IFIP International Conference on Artificial Intelligence Applications and Innovations. Cham: Springer Nature Switzerland, 2023.
13. Goel, Anubha, Amita Sharma, and Aparna Mehra. "Index tracking and enhanced indexing using mixed conditional value-at-risk." *Journal of Computational and Applied Mathematics* 335 (2018): 361-380.
14. Esfahanipour, Akbar, and Somayeh Mousavi. "A genetic programming model to generate risk-adjusted technical trading rules in stock markets." Expert Systems with Applications 38.7 (2011): 8438-8445.
15. Ansari, Yasmeen, et al. "A Multifaceted Approach to Stock Market Trading Using Reinforcement Learning." IEEE Access (2024).
16. Lim, Qing Yang Eddy, Qi Cao, and Chai Quek. "Dynamic portfolio rebalancing through reinforcement learning." Neural Computing and Applications 34.9 (2022): 7125-7139
17. AbdelKawy, Rasha, Walid M. Abdelmoez, and Amin Shoukry. "A synchronous deep reinforcement learning model for automated multi-stock trading." Progress in Artificial Intelligence 10.1 (2021): 83-97.
18. Kumlungmak, Kittiwin, and Peerapon Vateekul. "Multi-Agent Deep Reinforcement Learning With Progressive Negative Reward for Cryptocurrency Trading." *IEEE Access* 11 (2023): 66440-66455.
19. Haq, Anwar Ul, et al. "Forecasting daily stock trend using multi-filter feature selection and deep learning." *Expert Systems with Applications* 168 (2021): 114444.
20. Zhu, Wen, et al. "A data science framework for profit health assessment: development and validation." *Advances in Continuous and Discrete Models* 2024.1 (2024): 41.
21. Théate, Thibaut, and Damien Ernst. "An application of deep reinforcement learning to algorithmic trading." *Expert Systems with Applications* 173 (2021): 114632.
22. Carta, Salvatore, et al. "Multi-DQN: An ensemble of Deep Q-learning agents for stock market forecasting." Expert systems with applications 164 (2021): 113820.
23. Gao, Ruoyi, et al. "A Novel DenseNet-based Deep Reinforcement Framework for Portfolio Management." *2022 International Conference on Cyber-Enabled Distributed Computing and Knowledge Discovery (CyberC)*. IEEE, 2022.
24. Koratamaddi, Prahlad, et al. "Market sentiment-aware deep reinforcement learning approach for stock portfolio allocation." *Engineering Science and Technology, an International Journal* 24.4 (2021): 848-859.
25. Gašperov, Bruno, and Zvonko Kostanjčar. "Market making with signals through deep reinforcement learning." *IEEE access* 9 (2021): 61611-61622.
26. Charpentier, Arthur, Romuald Elie, and Carl Remlinger. "Reinforcement learning in economics and finance." *Computational Economics* (2021): 1-38.
27. De Moor, Bram J., Joren Gijsbrechts, and Robert N. Boute. "Reward shaping to improve the performance of deep reinforcement learning in perishable inventory management." *European Journal of Operational Research* 301.2 (2022): 535-545.
28. Andres, Alain, Esther Villar-Rodriguez, and Javier Del Ser. "Collaborative training of heterogeneous reinforcement learning agents in environments with sparse rewards: what and when to share?." *Neural Computing and Applications* 35.23 (2023): 16753-16780.

Received: August 10, 2025

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

29. Hirchoua, Badr, Brahim Ouhbi, and Bouchra Frikh. "Deep reinforcement learning based trading agents: Risk curiosity driven learning for financial rules-based policy." *Expert Systems with Applications* 170 (2021): 114553.
30. Pham, Uyen, Quoc Luu, and Hien Tran. "Multi-agent reinforcement learning approach for hedging portfolio problem." *Soft Computing* 25.12 (2021): 7877-7885.
31. Lussange, Johann, et al. "Modelling stock markets by multi-agent reinforcement learning." *Computational Economics* 57.1 (2021): 113-147.
32. Yu, Xiaoming, et al. "Dynamic stock-decision ensemble strategy based on deep reinforcement learning." *Applied Intelligence* 53.2 (2023): 2452-2470.
33. Gorde, P. S., & Borkar, S. N. (2024, August). Comparative Analysis of Linear Regression, Random Forest Regressor and LSTM for Stock Price Prediction. In 2024 8th International Conference on Computing, Communication, Control and Automation (ICCUBEA) (pp. 1-5). IEEE.
34. Yang, Steve Y., Yangyang Yu, and Saud Almahdi. "An investor sentiment reward-based trading system using Gaussian inverse reinforcement learning algorithm." Expert Systems with Applications 114 (2018): 388-401.
35. Park, Hyungjun, Min Kyu Sim, and Dong Gu Choi. "An intelligent financial portfolio trading strategy using deep Q-learning." Expert Systems with Applications 158 (2020): 113573.
36. Zhang, Huanming, Zhengyong Jiang, and Jionglong Su. "A deep deterministic policy gradient-based strategy for stocks portfolio management." 2021 IEEE 6th International Conference on Big Data Analytics (ICBDA). IEEE, 2021.
37. Borkar, S., & Jadhav, A. (2024). Reinforcement Learning Techniques for Stock Trading: A Survey of Current Research. Journal of Financial Data Science, 6(3).
38. Borkar, S. N., Jadhav, A., Dhablia, A., NandkishorAher, R., Aher, N. D., & Aware, A. A. (2023, August). Selection of Technical Indicators for Stock Market Prediction: Correlation Based Approach. In 2023 7th International Conference On Computing, Communication, Control And Automation (ICCUBEA) (pp. 1-6). IEEE.

Received: August 10, 2025