A Project Report On

# Deep Reinforcement Learning for Automated Stock Trading: An Ensemble Strategy

Under the supervision of

## Prof. Dr. Paresh Saxena

**SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS OF**

**CS F317: REINFORCEMENT LEARNING**

**BY**


| Sr No. | Name | ID Number |
|---|---|---|
| 1. | Harsh Vardhan Gupta | 2019B3A70630H |
| 2. | Aryan Kapadia | 2019B3A70412H |
| 3. | Parth Kulkarni | 2019B3A70706H |
| 4. | Shashwat Anand | 2019B3A70718H |
| 5. | Sai Srikar Chalamala | 2019AAPS0271H |
| 6. | Pranay Kumar Dasoju | 2019A7PS0006H |

**BIRLA INSTITUTE OF TECHNOLOGY AND SCIENCE PILANI**

**HYDERABAD CAMPUS**


**(16th April 2023)**

# Introduction

A crucial and difficult challenge for investors is learning how to make wise selections while trading stocks. There are 2 main important decisions that an investor needs to make:

1. Selection: Security selection implies picking individual stocks that the fund manager expects will outperform the market. Market timing implies betting on systematic risk factors.

2. Timing: Market timing is the strategy of making buying or selling decisions for financial assets by attempting to predict future market price movements.

The primary goal of this specific application of reinforcement learning is to assess whether a DRL agent can automatically make trading judgments and generate long-term consistent profits given that DRL has surpassed humans in many areas, such as playing Atari games. In this project, we will explore the ensemble trading strategy using three actor-critic-based algorithms: Proximal Policy Optimization (PPO), Advantage Actor-Critic (A2C), and Deep Deterministic Policy Gradient (DDPG). We will attempt to simulate the results of the primary paper and will try to suggest improvements to obtain better results.

# Related Work (Reviews)

1. Yang, H.. Deep Reinforcement Learning for Automated Stock Trading: An Ensemble Strategy, from https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3690996

   *Review:* This paper proposes an ensemble strategy that employs deep reinforcement schemes to learn a stock trading strategy by maximizing investment return. They trained a deep reinforcement learning agent and obtained an ensemble trading strategy using three actor critic-based algorithms: Proximal Policy Optimization (PPO), Advantage Actor-Critic (A2C), and Deep Deterministic Policy Gradient (DDPG). The proposed deep ensemble strategy has outperformed all three individual algorithms, the Dow Jones Industrial Average and min-variance portfolio allocation. Initially, an Environment was built, and then secondly, the three algorithms were trained, and lastly, the three agents were ensembled using the Sharpe ratio. A higher Sharpe ratio verifies the effectiveness.

2. L. Chen and Q. Gao, "Application of Deep Reinforcement Learning on Automated Stock Trading," 2019 IEEE 10th International Conference on Software Engineering and Service Science (ICSESS), Beijing, China, 2019, pp. 29-33, doi:10.1109/ICSESS47205.2019.9040728, from https://ieeexplore.ieee.org/abstract/document/9040728?casa_token=xAs-JBgbJhQAAAAA:cUQFlgLHauvepFTGhJZW4y8DDWQJwRebP4I9TabsHwrJqpqVCcw2DBYiOE0nqbPi4xjh-OtCxVuf_Go

   *Review:* This paper tries to solve the agent that can automatically make stock trading decisions to achieve long-term stable profits by applying Deep Q-Network (DQN) and Deep Recurrent Q-Network (DRQN). The S&P500 ETF is selected as the trading asset. Furthermore, the agent's performance is evaluated by comparing it with benchmarks of Buy and Hold (BH) and Random action-selected DQN. The results showed that the DRQN trader is even better than the DQN trader mainly because the recurrence framework can discover and exploit patterns hidden in time-related sequences.

3. AbdelKawy, R. (2021, January 5). A Synchronous Deep Reinforcement Learning Model for Automated Multi-Stock Trading. SpringerLink, from https://link.springer.com/article/10.1007/s13748-020-00225-z?error=cookies_not_supported&code=a004a2ce-bee9-464c-a332-65d488ce6e6a

   *Review:* This paper presents a novel multi-stock trading model based on free-model synchronous multi-agent deep reinforcement learning, which interacts with the market and captures the financial market dynamics. They used two types of deep neural networks, the Deep Belief Network (DBN) and the long short-term memory (LSTM) network. The DBN is a feature extraction network, while the LSTM network is suitable for predicting long-time series data. The proposed network has improved the performance of policy-based and value-based reinforcement learning techniques compared to other existing models.

4. W. Si, J. Li, P.Ding and R. Rao, "A Multi-objective Deep Reinforcement Learning Approach for Stock Index Future's Intraday Trading," 2017 10th International Symposium on Computational Intelligence and Design (ISCID), Hangzhou, 2017, pp. 431-436, doi: 10.1109/ISCID.2017.210, from https://ieeexplore.ieee.org/abstract/document/8283307?casa_token=mvJWTD9yr_wAAAAA:hqNVbBRLBG8fV5XNe9C7ZATQkCOYBelsZPi0KTd1B-dkIodE6jm8ncnIp4jiDGprRvjBWjspFf-GT8g

   *Review:* This paper presents a novel approach for using DRL to optimize trading decisions in the stock index futures market. The paper begins by discussing the challenges of intraday

trading in the stock index futures market, including the high volatility and uncertainty of the market. The authors propose a multi-objective DRL framework, which aims to simultaneously optimize multiple objectives, including maximizing returns and minimizing risk. The results of the study demonstrate that the proposed multi-objective DRL approach is able to generate profitable trading strategies while also effectively managing risk. The approach outperforms several benchmark methods, including a random policy and a buy-and-hold strategy. The authors also compare the performance of the multi-objective DRL approach to that of a single-objective DRL approach and show that the multi-objective approach is able to achieve superior results.

5. Liu, X. (2021, November 4). FinRL: Deep Reinforcement Learning Framework to Automate Trading in Quantitative Finance, from https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3955949

   *Review:* This paper discusses the potential benefits of using DRL to address the challenges of automated trading in quantitative finance. A DRL framework called FinRL is proposed, which is specifically designed to address the unique characteristics of financial data, such as high dimensionality, non-stationarity, and noisy signals. The FinRL framework consists of several components, including a trading environment, a deep neural network to represent the trading policy, a replay buffer to store past experiences, and a DRL algorithm to update the policy. The paper describes the experimental setup, which involves training the FinRL framework on historical market data and evaluating its performance on a test dataset. The results of the study demonstrate that the FinRL framework can generate profitable trading strategies across a range of financial assets, including stocks, currencies, and futures. The approach outperforms several benchmark methods, including a buy-and-hold strategy and a technical analysis-based strategy. The authors also compare the performance of the FinRL framework to that of several other DRL-based trading strategies and show that FinRL is able to achieve superior results.

6. Kabbani, T., & Duman, E. (2022). Deep Reinforcement Learning Approach for Trading Automation in the Stock Market. IEEE Access, 10, 93564-93574, from https://arxiv.org/pdf/2208.07165.pdf

   *Review*: This paper aims to investigate the performance of a DRL-based trading strategy on the stock market. They proposed a DRL-based trading strategy that uses a combination of a deep neural network and a Q-learning algorithm. The neural network is trained on historical stock market data to predict future stock prices. The Q-learning algorithm is used to learn the optimal trading policy based on the predicted stock prices. The results showed that the DRL-based strategy outperformed both the buy-and-hold strategy and the traditional strategy in terms of cumulative returns. Additionally, the DRL-based strategy had a lower risk than the traditional strategy, as measured by the maximum drawdown.

# Dataset and Reference Paper Results

This report uses the DOW Jones Industrial Average, which is the stock market index of the 30 most prominent companies listed in the stock exchanges in the United States. It is one of the oldest and most commonly followed stock indexes. The training dates are from 1st Jan 2010 to 1st October 2021 and the testing dates are from 1st October 2021 to 1st March 2023.

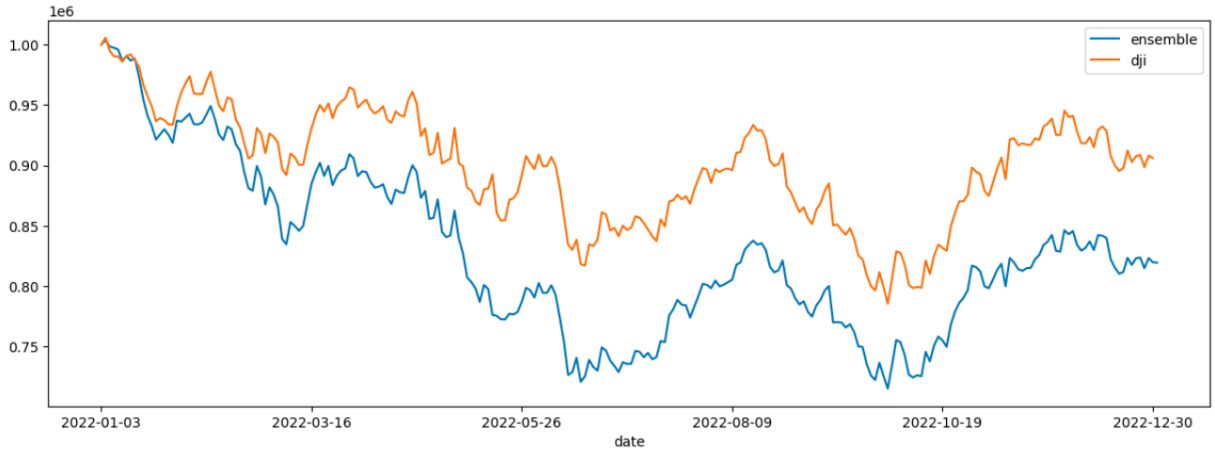
The 30 companies are: American Express Co (AXP), Amgen Inc (AMGN), Apple Inc (AAPL), Boeing Company (BA), Caterpillar Inc (CAT), Cisco Systems Inc (CSCO), Chevron Corporation (CVX), Goldman Sachs Group Inc (GS), Home Depot Inc (HD), Honeywell International Inc (HON), IBM (IBM), Intel Corporation (INTC), Johnson & Johnson (JNJ), Coca-Cola Company (KO), JPMorgan Chase & Co (JPM), McDonald's Corporation (MCD), 3M Company (MMM), Merck & Co Inc (MRK), Microsoft Corporation (MSFT), Nike Inc (NKE), Procter & Gamble Company (PG), Travelers Companies Inc (TRV), UnitedHealth Group Inc (UNH), Salesforce.com Inc (CRM), Verizon Communications Inc (VZ), Visa Inc (V), Walgreens Boots Alliance Inc (WBA), Walmart Inc (WMT), Walt Disney Company (DIS), Dow Inc (DOW)

## About Environment:

The stock dimension is 30, and the state space is 181. For a single stock, the action space is defined as $\{-k, ..., -1, 0, 1, ..., k\}$, where k and $-k$ present the number of shares we can buy and sell, and $k \leq hmax$ where $h_{max}$ is a predefined parameter that sets as the maximum number of shares for each buying action. Therefore, the action space is $(2k+1)^{30}$.

| Iteration | Val_Start | Val_End | Model Used | A2C Sharpe | PPO Sharpe | DDPG Sharpe |
|---|---|---|---|---|---|---|
| 126 | 2021-10-04 | 2022-01-03 | DDPG | 0.0778 | 0.0675 | 0.1535 |
| 189 | 2022-01-03 | 2022-04-04 | A2C | -0.1432 | -0.2266 | -0.2069 |
| 252 | 2022-04-04 | 2022-07-06 | DDPG | -0.2113 | -0.2509 | -0.1596 |
| 315 | 2022-07-06 | 2022-10-04 | DDPG | -0.1607 | -0.2115 | -0.1212 |

The results for the model suggested by the reference paper on DOW JONES index are shown in the above table. From 3rd Jan 2022 to 4th April 2022, the model used is A2C. For all other dates in the testing set, the model used is DDPG. The model selected for a particular period is the one that has the highest Sharpe ratio among the three models.




In the above image, on the y-axis, we have the amount invested (in the units of million USD), and on the x-axis, we have the date. The initial portfolio value is set to $1 million. The orange line represents the portfolio value when invested in Dow Jones Index. In contrast, the blue line represents the portfolio value using the strategy suggested by the ensemble model. It can be observed that the portfolio value, as indicated by the model, follows the trends of the DOW JONES index, but the magnitude is undervalued, in general.

# Experiments and Improvements

We attempted to use a similar ensemble model for Indian markets. We used the top 30 companies belonging to the famous "NIFTY FIFTY" index. The model was trained from 1st Jan 2005 to 1st Oct 2021 and then tested from 2nd Oct 2021 to 1st April 2023. Using the default as suggested in the paper, we trained the RL model and obtained the following results:

| Iteration | Val_Start | Val_End | Model Used | A2C Sharpe | PPO Sharpe | DDPG Sharpe |
|---|---|---|---|---|---|---|
| 126 | 2021-10-04 | 2022-01-04 | PPO | -0.1883 | **0.1753** | 0.0029 |
| 189 | 2022-01-04 | 2022-04-06 | PPO | -0.1856 | **0.1193** | -0.2133 |
| 252 | 2022-04-06 | 2022-07-07 | A2C | **-0.0085** | -0.4177 | -0.0440 |
| 315 | 2022-07-07 | 2022-10-10 | A2C | **-0.4185** | -0.0430 | 0.3786 |

The results for the model suggest that from 4th Oct 2021 to 6th Apr 2022, the best model was PPO, while for the remaining dates in the testing set, the best model is A2C. Comparing the

portfolio values generated by the model against the NIFTY portfolio in the graph below it can be observed that the ensemble strategy performs poorly. The generated portfolio value, although following the trend, is significantly lower than the NIFTY values. The time taken to train the model was: **38.133 minutes**

As a result, some changes were made in the model, which are listed below:


## 1. Training on a higher number of time steps:

The default model of the paper ran all three approaches, i.e., A2C, PPO, and DDPG on 10,000-time steps. We tried increasing this to 30,000 so that the agent could learn the strategies better on each episode. The results of the same are mentioned below:

| Iteration | Val_Start | Val_End | Model Used | A2C Sharpe | PPO Sharpe | DDPG Sharpe |
|---|---|---|---|---|---|---|
| 126 | 2021-10-04 | 2022-01-04 | PPO | 0.1074 | **0.1547** | 0.1354 |
| 189 | 2022-01-04 | 2022-04-06 | PPO | -0.1171 | **0.075292** | -0.2133 |
| 252 | 2022-04-06 | 2022-07-07 | A2C | **-0.1844** | -0.2609 | -0.2296 |
| 315 | 2022-07-07 | 2022-10-10 | A2C | **0.4420** | 0.2884 | 0.0905 |

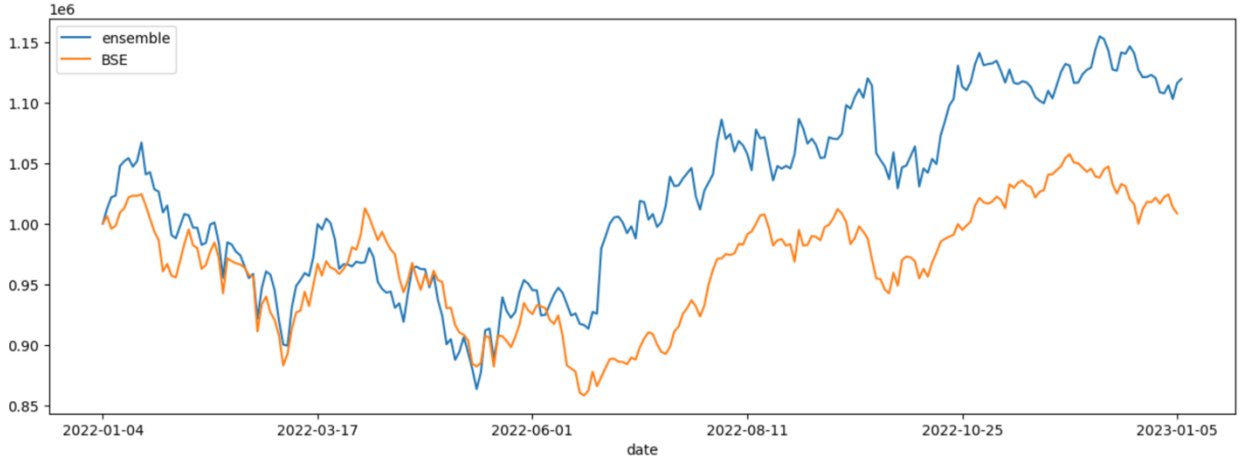
The ensemble strategy chosen by the model is same as that suggested by the previous model. However, the Sharpe ratio for the period 7th July 2022 to 10th October 20, 2022, has **significantly** improved from -0.4185 to 0.4420 which represents the portfolio's excess returns adjusted for the given level of risk. This can also be observed in the graph below where we can see that the value of the ensembled portfolio has significantly improved over the previous model in addition to the general trends of the NIFTY portfolio being followed. The time taken to train the model was: **92.638 minutes**.


## 2. Changing the optimizer and Discount Factor:

The default model used RMS Prop as the DL optimized with the discount factor of 0.1. We tried using ADAM optimizer, which combines RMS Prop with Momentum, to achieve faster convergence. The discount factor was set to 0.2 to give more weightage for future rewards, and the number of time steps is set to a default of 10000 steps. The results are shown below:

| Iteration | Val_Start | Val_End | Model Used | A2C Sharpe | PPO Sharpe | DDPG Sharpe |
|---|---|---|---|---|---|---|
| 126 | 2021-10-04 | 2022-01-04 | A2C | **0.2722** | -0.0012 | 0.1269 |
| 189 | 2022-01-04 | 2022-04-06 | DDPG | 0.0210 | -0.1522 | **0.0429** |
| 252 | 2022-04-06 | 2022-07-07 | A2C | **0.2147** | -0.0557 | 0.1114 |
| 315 | 2022-07-07 | 2022-10-10 | DDPG | 0.2975 | 0.1990 | **0.3893** |

The agent chose A2C model during the 1st and 3rd periods while DDPG was chosen for 2nd and 4th periods as shown in the table above. By far, this model has the best performance, and it **outperforms** the NIFTY portfolio by a significant amount. The Sharpe ratio increased significantly in almost all the periods. The comparison can be seen in the image below. The time taken to train the model was: **35.9758 minutes** which is around **12.5%** faster than the base model that has RMS Prop as the optimizer.




## 3. Improving the Previous model:

To improve the model further we tried to increase the discount factor to 0.5 in order to provide more weightage to future rewards. Also, the time steps were increased to 30000 and again the ADAM optimizer was used. The results are shown below:

| Iteration | Val_Start | Val_End | Model Used | A2C Sharpe | PPO Sharpe | DDPG Sharpe |
|---|---|---|---|---|---|---|
| 126 | 2021-10-04 | 2022-01-04 | A2C | **-0.0249** | -0.3637 | -0.0669 |
| 189 | 2022-01-04 | 2022-04-06 | DDPG | -0.0829 | -0.0930 | **-0.0578** |
| 252 | 2022-04-06 | 2022-07-07 | DDPG | -0.0836 | -0.0618 | **0.0383** |
| 315 | 2022-07-07 | 2022-10-10 | DDPG | 0.0691 | 0.0409 | **0.1824** |

The model selects A2C model for the first period, and then selects DDPG approach for the rest three periods. This model clearly outperformed the previous model, as seen in the graph below. The portfolio value at the end of the test period is about $1.2 million as compared to the previous model which achieved the final portfolio value of about $1.15 million.


# Conclusion

This report explores the ensemble strategy consisting of A2C, PPO, and DDPG suggested in the main reference paper on US and Indian market data. Further it can be concluded that ensemble strategy is quite efficient but requires certain extent of fine tuning of parameters such as optimizers, time steps, discount factor, etc. to outperform the market. However, markets are subjected to various factors such as Market volatility, Data quality and availably, stock splits, External factors such interest rate changes, geopolitical events and trading costs and fees. As a result using this strategy in practical purposes is not always feasible.

# Contributions

| Sr No. | Name | Contribution |
|---|---|---|
| 1. | Harsh Vardhan Gupta | Code, Improvements |
| 2. | Aryan Kapadia | Code, Improvements |
| 3. | Parth Kulkarni | Report, Observations |
| 4. | Shashwat Anand | Report, Observations |
| 5. | Sai Srikar Chalamala | Report, Literature Review |
| 6. | Pranay Kumar Dasoju | Report, Literature Review |