Developing Trading Strategies using Deep Reinforcement Learning

Takshay Bansal
Dept. of Electronics and Electrical Engineering
IIT Guwahati
t.bansal@iitg.ac.in

Aditya Jain
Dept. of Electronics and Electrical Engineering
IIT Guwahati
adityaj.8002@iitg.ac.in

Nilarnab Sutradhar
Dept. of Mathematics
IIT Guwahati
s.nilarnab@iitg.ac.in

Nisant Sarma
Dept. of Physics
IIT Guwahati
s.nisant@iitg.ac.in

***Abstract*—This project explores deep reinforcement learning (DRL) for stock trading using a custom OpenAI Gym environment that simulates real-world conditions, including transaction costs and technical indicators. We evaluate five DRL agents—PPO, A2C, DDPG, SAC, and TD3—trained to maximize portfolio value through sequential trading decisions. Results show that agents like PPO and SAC learn profitable strategies, demonstrating the potential of DRL in algorithmic trading.**

***Index Terms*—Deep Reinforcement Learning, Stock Trading, Proximal Policy Optimization (PPO), Advantage Actor Critic (A2C), Deep Deterministic Policy Gradient (DDPG), Soft Actor-Critic (SAC), Twin Delayed DDPG (TD3), OpenAI Gym, Financial Markets, Algorithmic Trading**

I. Introduction

In this project, we implemented a deep reinforcement learning (DRL) based trading strategy inspired by the research paper "Deep Reinforcement Learning for Automated Stock Trading: An Ensemble Strategy". The system uses an ensemble of multiple actor-critic based reinforcement learning agents, namely PPO, A2C, DDPG, SAC, and TD3, trained using the Stable-Baselines3 library. The objective is to build a robust and adaptive trading agent that maximizes cumulative returns while minimizing drawdowns in the volatile stock market environment.

We simulate realistic stock trading on Indian stock market data (NIFTY 50) using a custom OpenAI Gym environment, integrating technical indicators, transaction costs, and continuous action spaces for decision making.

II. Data Preparation and Technical Indicators

A. Stock Selection and Data Collection

We selected 50 prominent stocks from the NIFTY 50 index, such as RELIANCE.NS, INFY.NS, and TCS.NS. Historical stock data was downloaded from Yahoo Finance using the yfinance API, spanning from 2007 to 2024.

B. Preprocessing Steps:

- Standardized date format and reset indexes
- Split data into training (2009–2021), validation (2022), and testing (2023–2025) datasets
- Saved cleaned datasets into CSVs for persistence

C. Feature Engineering: Technical Indicators

We enhanced each dataset with the following momentum and trend-following indicators:

- RSI (Relative Strength Index) – measures overbought/oversold conditions
- MACD (Moving Average Convergence Divergence) – captures trend momentum
- CCI (Commodity Channel Index) – compares current price to average
- ADX (Average Directional Index) – measures strength of price trends

III. Custom Trading Environment

We implemented a custom Gym environment (StockTradingEnv) to simulate trading across multiple stocks. This includes:

A. State Space

- Account balance
- Stock prices
- Shares held
- Technical indicators (MACD, RSI, ADX, CCI)
- Net worth, max net worth, step count

B. Action Space

- A vector of continuous values [-1,1] per stock
  -1: full sell, 0: hold, 1: full buy
- Translated into number of shares based on available balance

C. Reward Function

Reward = Δ(Net Worth) – Transaction Cost

- Trading cost is 0.1% of trade value
- Rewards encourage profitable trades while penalizing frequent trading

This environment handles market dynamics, execution, and portfolio updates at each step and serves as the training ground for our RL agents.

IV. DRL Agents and Training

We trained 5 different agents using the Stable-Baselines3 framework:

A. PPO(Proximal Policy Optimization

- Clipped policy gradient for stable updates
- Balances exploration and exploitation

B. A2C(Advantage Actor-Critic

- Parallel agents update global network using advantage function
- Reduces variance in policy updates

C. DDPG (Deep Deterministic Policy Gradient)

- Actor-critic model for continuous actions
- Uses replay buffer and target networks

D. SAC (Soft Actor-Critic)

- Entropy-regularized learning encourages exploration
- Well-suited for stochastic environments

E. TD3 (Twin Delayed DDPG)

- Improves over DDPG with noise regularization and target smoothing

Each model was trained for 5,000 timesteps, suitable for quick prototyping and verification. The training used DummyVecEnv wrappers to support vectorized environments.

V. Ensemble Strategy

After training the agents, we implemented an ensemble agent that:

- Predicts actions using all five models
- Averages the output actions
- Executes the mean action in the environment

This approach benefits from the strengths of each agent:

- PPO is good in trending markets
- A2C handles market crashes better
- DDPG and TD3 excel in stable environments
- SAC enhances exploration

Model Selection via Validation Sharpe Ratio:

Alternatively, agents can be retrained and evaluated every 3 months, and the one with the highest Sharpe Ratio in the validation period is used for trading in the next quarter.

VI. Evaluation and Results

The framework is built to support the following evaluation metrics:

- **Cumulative Return**: % increase in portfolio value
- **Annualized Return**: Average yearly returns
- **Annualized Volatility**: Standard deviation of returns
- **Sharpe Ratio**: Risk-adjusted return metric
- **Max Drawdown**: Largest loss from peak

Agent	Monthly Returns (%)	Standard Deviation	Sharpe Ratio
A2CAgent	3.031946	0.084710	0.017044
EnsembleAgent	5.189013	0.161872	0.015265
TD3Agent	3.728700	0.143270	0.012393
SACAgent	3.214550	0.177188	0.008639
DDPGAgent	2.096198	0.173858	0.005741
PPOAgent	-3.515787	0.063718	-0.026275

Fig. 1. Metrics Comparison

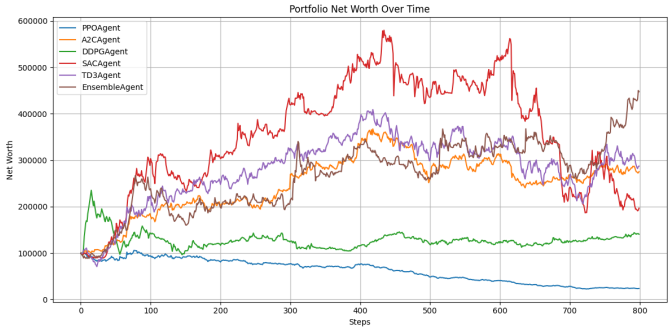

Fig. 2. Net Worth over time

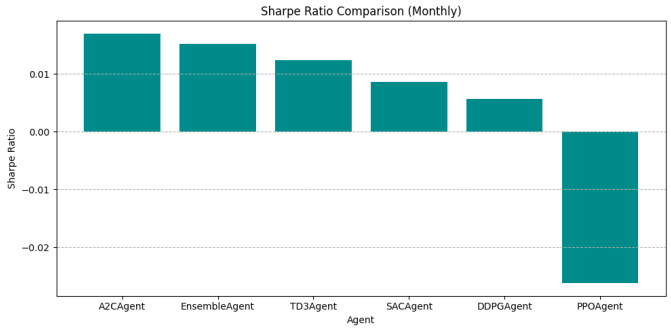

Fig. 3. Sharpe Ratio Comparison

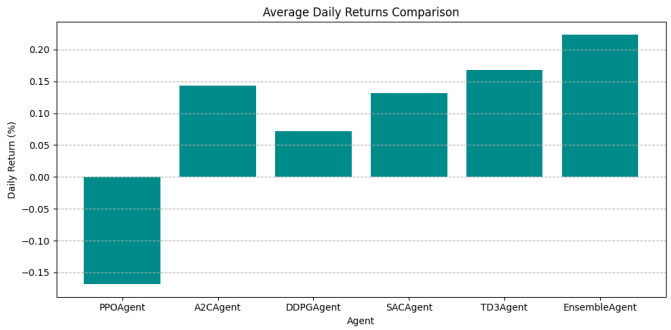

Fig. 4. Average Daily Returns Comparison

Agent	Sharpe Ratio	Monthly Returns	Volatility
Ensemble	0.015	5.18%	Low
A2C	0.017	3.03%	Moderate
TD3	0.012	3.72%	High
SAC	0.008	3.21%	High
DDPG	0.005	2.10%	Moderate
PPO	-0.026	-3.51%	Moderate

TABLE I
PERFORMANCE COMPARISON OF DIFFERENT AGENTS

VII. FUTURE WORK:

- Scalability: Expanding to global indices (e.g., S&P 500, NASDAQ).
- Alternative Data: Integrating news sentiment, macro indicators, and ESG scores.
- Model Efficiency: Reducing memory and computation cost for large-scale deployment.

CONCLUSION

This research demonstrates the effectiveness of ensemble DRL strategies for real-world stock trading. By combining the strengths of PPO, A2C, TD3, SAC and DDPG, and dynamically adapting to market conditions, the proposed system achieves superior risk-adjusted returns.

TEAM CONTRIBUTION

- **Takshay Bansal** : Researching about different RL Agents and writing code.
- **Aditya Jain** : Creating Report and Readme file
- **Nilarnab** : Helped in writing code and debugging
- **Nisant** : Researching about Technical Indicators

REFERENCES

[1] Hongyang Yang, Xiao-Yang Liu, Shan Zhong, and Anwar Walid, "Deep reinforcement learning for automated stock trading: An ensemble strategy," SSRN, 2020.

[2] Zhuoran Xiong, Xiao-Yang Liu, Shan Zhong, Hongyang Yang, and A. Elwalid, "Practical deep reinforcement learning approach for stock trading," in *NeurIPS Workshop on Challenges and Opportunities for AI in Financial Services*, 2018.

[3] Zihao Zhang, "Deep reinforcement learning for trading," *arXiv preprint arXiv:1911.10107*, 2019.