Deep Reinforcement Learning for Automated Stock Trading

組別二

113352030 金碩一胡益鳴
113352032 金碩一翁晟睿
113352019 金碩一黃得晉
114352018 金碩一楊成豐
114363013 企碩一李姵琪

摘要（翁晟睿編輯）

本研究主要參考 Yang et al. (2020) 中的方法，使用 A2C（Advantage Actor Critic）、PPO（Proximal Policy Optimization）、DDPG（Deep Deterministic Policy Gradient）三種強化學習（Reinforcement Learning, RL）演算法，訓練三個以最大化股票投資報酬率為目標的代理人，並組成一個集成的股票交易策略。該集成策略整合了三種演算法的優點，從而能夠穩健地適應不同的市場情況。

除此之外，本研究也對 Yang et al. (2020) 中的方法進行了一些優化，例如：在訓練代理人的過程中，環境變數中加入更多種類的技術指標，並利用隨機森林篩選出重要性最高的前五個技術指標；此外，在風險控制的門檻上進行調整，幫助代理人在交易時更能及早發現熊市的出現並做出應對。

本研究與 Yang et al. (2020) 相同，皆使用具有充足流動性的道瓊工業指數中的 30 檔成分股作為投資標的並進行回朔測試，結果顯示本研究的策略在報酬率、風險控制與最大回檔比率等方面皆表現優於原始研究。

壹、簡介及文獻探討(黃得晉編輯)

在近年金融科技（FinTech）迅速發展的浪潮下，強化學習（Reinforcement Learning, RL）作為一種具備自我學習與策略優化能力的演算法架構，逐漸受到金融投資領域的高度重視。相較於傳統量化策略需依賴專家定義特徵與規則，強化學習能透過與市場環境互動不斷修正行動策略，展現更強的適應性與決策彈性。

Yang et al. (2020) 提出一套以深度強化學習為基礎的股票交易架構，結合三種不同特性的演算法——A2C（Advantage Actor Critic）、PPO（Proximal Policy Optimization）與 DDPG（Deep Deterministic Policy Gradient）——並採用集成策略（Ensemble Strategy）融合三者優勢，於道瓊工業指數成分股上進行實證測試，展現出優於指數的累積報酬與風險控管能力。

本研究在其基礎上進一步擴充並優化整體策略架構，主要從三個方向切入：第一，透過特徵工程引入多種技術指標並採用隨機森林（Random Forest）進行篩選，提升狀態資訊純度與模型效率；第二，設計符合 OpenAI Gym 框架的多資產環境，模擬實際交易限制；第三，提出動態選模（Dynamic Model Selection）與風險控制機制（如 Turbulence Index）以提升策略穩健性與適應性。

本研究希望藉由整合深度強化學習、特徵篩選、滾動視窗與風險管理等模組，建構一套具備自我調適與長期穩定性的股票交易系統，並於歷史數據中驗證其超越市場報酬與控制風險的潛力。

貳、資料(胡益鳴編輯)

(一) 研究標的與資料來源

- 標的資產：道瓊工業指數 30 檔成分股（Dow 30）
- 期間範圍：2009/01/01 ～2024/10/07（共約 4000 個交易日）
- 資料頻率：日資料（Daily）
- 原始欄位：tic、open、high、low、adjcp、volume
- 資料來源：WRDS

參、研究方法

一、深度強化學習 (RL) 代理人訓練

(一) Actor–Critic 架構（翁晟睿編輯）

本研究使用 A2C、PPO、DDPG 三種強化學習演算法，分別訓練三個以最大化股票報酬為目標的代理人，並組成一個集成的交易策略。這三種代理人皆使用類似的 Actor–Critic 架構。

在此架構下，**Actor** 是負責決策的部分，會根據當前的環境狀態 S_t，選擇對應的動作 a_t（如買進、賣出或持有）。動作執行後，環境會進入下一個狀態 S_{t+1}。

另一方面，**Critic** 負責評估這個動作的價值。它根據當前狀態 S_t 和下個狀態 S_{t+1}，透過 優勢函數（Advantage Function）計算該動作實際獲得的報酬與原先預期的差距。

這個差距，亦即 Advantage 值，會被納入 Actor 的損失函數，幫助 Actor 調整策略，使其更傾向於選擇具有高優勢的動作。Critic 自身也會透過損失函數（如 MSE）來提升其對報酬的預測能力。

三種代理人的差異主要體現在 Actor 和 Critic 的更新方式與損失函數形式的不同。

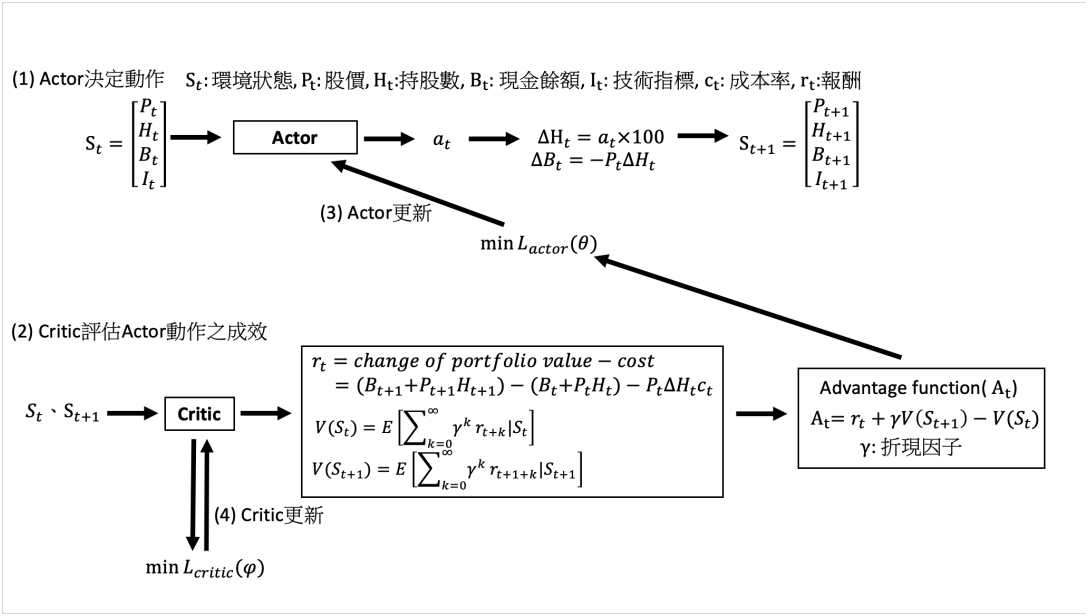

圖 1: Actor-Critic 架構圖：Actor 根據環境狀態選擇動作，Critic 評估並回饋優勢函數

(二) A2C（翁晟睿編輯）

A2C 是本研究中用於訓練 RL 代理人的一種演算法，其一樣有前一節所介紹的 Actor–Critic 架構，但其中 Actor 是利用一常態分布，並從中隨機抽樣以決定要採取的動作。假設抽樣結果為 a_t，若 $a_t > 0$ 則買進股票，若 $a_t < 0$ 則賣出股票，若 $a_t = 0$ 則繼續持有原有部位。Critic 則會在得知 S_t 和 S_{t+1} 後，計算 Actor 執行動作後的確切報酬 r_t，以及 Actor 執行動作前後的長期預期報酬 $V(S_t)$ 與 $V(S_{t+1})$，並進一步計算優勢函數（Advantage Function）A_t。

A_t 可拆成兩部分來解讀：第一部分為 $r_t + \gamma V(S_{t+1})$，即 Actor 執行動作後當期的報酬加上未來預期報酬的折現值，而第二部分為 $V(S_t)$，即 Actor 執行動作前的長期預期報酬，兩部分相減即為 A_t，其含義為 Actor 所執行的動作比預期情況好多少，即為 Actor 所執行的動作的優勢。

如前 Actor–Critic 架構所介紹，Critic 可以透過計算 A_t 以評估 Actor 所執行動作的優勢，並透過將 A_t 包含在 Actor 的目標式中協助其更新動作分配的參數。

A2C 更新 Actor 的方式是通過極大化目標式 (1)，其中的 GAE（Generalized Advantage Estimation）為一種 A 的形式，差別在於 GAE 相比 A_t 多考慮了長期的優勢，

而即為動作的分配函數，因此 Actor 的目標可以看成極大化某動作產生的優勢乘以該動作發生的機率：

$$\max J(\mu, \sigma^2) = \mathbb{E}\left[\sum_{t=1}^{T} \mathrm{GAE}_t \cdot \ln f(a_t \mid \theta, S_t)\right], \quad f \sim \mathcal{N}(\mu, \sigma^2) \tag{1}$$

其中：

$$\mathrm{GAE}_t = \sum_{l=0}^{\infty} (\gamma\lambda)^l A_{t+l}, \quad A_t = r_t + \gamma V(S_{t+1}) - V(S_t)$$

γ：折現因子，λ：超參數。

A2C 更新 Critic 的方式則是通過極小化目標式 (2)，該目標式又稱 TD target（Temporal Difference Target），為一種 MSE 的形式，旨在增加 Critic 對於 $V(S_t)$ 的預測能力，其中 φ 為透過神經網路訓練 Critic 預測 $V(S_t)$ 時的參數，如權重和誤差，而增加 $V(S_t)$ 的預測能力亦能同時增加 Critic 預測 A_t 的能力，以幫助 Actor 更精確地更新其動作分配。

$$\min \mathcal{L}_{\text{critic}}(\varphi) = [r_t + \gamma V(S_{t+1}) - V(S_t)]^2 \tag{2}$$

(三) PPO（黃得晉編輯）

承接前一節對 A2C 的介紹，本節聚焦說明其改良版 Proximal Policy Optimization（以下簡稱為 PPO）。PPO 同樣採用 Actor–Critic 架構，將策略網路（Actor）與價值網路（Critic）分開訓練。Actor 根據觀測到的狀態 S_t 輸出動作的分佈參數，例如在連續動作空間下生成常態分布 $\mathcal{N}(\mu, \sigma)$；Critic 則負責估算該狀態下的價值 $V(S_t)$，作為 TD target 與 Advantage function 的參考基準。

PPO 最大的特色在於其「鄰近策略更新（Proximal Policy Update）」的理念，目的是在每次更新策略時，限制新策略與舊策略之間的差異幅度，避免策略偏移過大而導致學習不穩。這個限制是透過 Clipped Objective 損失函數來實現的。訓練流程如下：

首先，使用舊策略 $\pi_{\theta^{\text{old}}}$ 與環境互動，收集資料集 $(s_t, a_t, r_t, \log \pi_{\theta^{\text{old}}}(a_t \mid s_t), V_\varphi(s_t))$，再使用 GAE（Generalized Advantage Estimation）計算每筆樣本對應的 Advantage 值 A_t。

接著將資料隨機打亂成 mini-batch，並重複多輪（例如四輪）epoch 更新，極大化樣本資訊的使用效率。

策略梯度部分，PPO 首先計算新舊策略的機率比值：

$$r_t = \frac{\pi_\theta(a_t \mid s_t)}{\pi_{\theta^{\text{old}}}(a_t \mid s_t)}$$

若直接極大化 $r_t A_t$，當比值偏離太多時容易導致策略不穩，因此 PPO 採用以下裁

剪形式的損失函數來限制更新範圍：

$$\mathcal{L}_{\text{clip}}(\theta) = \mathbb{E}_t \left[\min \left(r_t A_t, \text{clip}(r_t, 1-\varepsilon, 1+\varepsilon) A_t \right) \right]$$

其中 ε 為超參數，我們設為 0.1。

此外，PPO 的完整損失函數還包含 Critic 的均方誤差：

$$\mathcal{L}_{\text{critic}} = c_v \cdot (V(s_t) - G_t)^2$$

以及策略的熵獎勵項：

$$\mathcal{L}_{\text{entropy}} = -c_e \cdot \mathcal{H}(\pi_\theta)$$

三者加總成總體損失，對 Actor 與 Critic 進行聯合優化。

一次外層迭代的整體流程如下：

1. 使用舊策略收集一批交互軌跡（trajectory）
2. 利用 GAE 計算 Advantage 並標準化
3. 將資料拆成 mini-batch 並進行多輪 epoch 更新
4. 使用剪裁後的損失函數對 Actor 和 Critic 同步更新
5. 檢查平均報酬、策略熵或 KL 散度是否達到提前停止標準

與 A2C 相比，PPO 優點在於樣本利用率更高，穩定性更佳，且對策略變動有明確控制；缺點則是訓練成本較高，需進行更多次內部 epoch 迭代，且需微調超參數如 ε、epoch 數量、mini-batch 大小等。

在實務操作上，PPO 經常用於策略表現優化階段，通常會先用 A2C 驗證可行性，再切換 PPO 進行策略精煉。此外亦可加入自適應 KL 散度門檻，或透過貝氏優化進一步調整裁剪係數與學習速率，以平衡穩定性與收斂效率。

(四) DDPG（楊成豐編輯）

本節聚焦介紹 DDPG（Deep Deterministic Policy Gradient），這是一種專為連續動作空間設計的 off-policy 強化學習演算法，結合了 DQN 的穩定訓練技巧與 Actor–Critic 架構的策略學習能力。相較於 A2C 與 PPO 這類 on-policy 策略，DDPG 在樣本利用與連續控制表現上展現出明顯優勢，適用於像是資產配置這種精細控制問題。

DDPG 使用兩組獨立的 Actor 與 Critic 網路，其中：

- Actor：輸出確定性動作 $a = \mu(s)$，代表每個資產的配置比例。
- Critic：接收 (s, a) 配對，估計對應的 Q 值 $Q(s, a)$，評估該動作的長期價值。

不同於 PPO 輸出動作分布並進行抽樣，DDPG 直接產出確定性動作，因此需額外引入探索雜訊（如 Ornstein-Uhlenbeck 過程）來促進探索。

為了提升學習穩定性與效率，DDPG 結合以下三個關鍵模組：

- **Replay Buffer**：將歷史交互資料儲存起來，打亂順序後重複抽樣使用，大幅提升樣本利用效率並避免時序相關性。
- **Target Network**：為 Actor 與 Critic 各維護一套目標網路，以 soft update 的方式緩慢同步參數，減少 Q 值估計的不穩定。
- **探索雜訊（Exploration Noise）**：將時間相關的 OU 雜訊加入動作，有助於模擬市場慣性波動，提升早期探索能力。

訓練流程簡述：

1. Actor 使用當前策略（加上探索雜訊）與環境互動，儲存經驗至 replay buffer。
2. 當 buffer 達一定大小，隨機抽樣一批資料，分別更新：
   - **Critic**：使用 TD target 計算誤差並進行 MSE 回歸，公式為：

$$y_t = r_t + \gamma Q_{\phi'}(s_{t+1}, \mu_{\theta'}(s_{t+1}))$$

   - **Actor**：最大化 Critic 對自身動作的評價，即最小化：

$$\mathcal{L}_{\text{actor}} = -Q_{\phi}(s, \mu_{\theta}(s))$$

3. 使用 soft update 更新 target networks：

$$\theta' \leftarrow \tau\theta + (1-\tau)\theta', \quad \phi' \leftarrow \tau\phi + (1-\tau)\phi'$$

DDPG 使用 MLP 作為基礎結構，Actor 與 Critic 分別接受狀態（或狀態 + 動作）為輸入，經過兩層 64-unit 隱藏層後輸出連續動作或 Q 值。此結構在金融交易中能精細控制每筆資產的買入與賣出比例，特別適合需要連續調節的場景。

(五) A2C、PPO、DDPG 之比較（楊成豐編輯）

下表為三種強化學習演算法的重點比較摘要：

項目	A2C	PPO	DDPG
策略類型	隨機策略（Stochastic）	隨機策略（Stochastic）	確定性策略（Deterministic）
動作空間支援	離散／連續	離散／連續	僅連續
資料使用	單次使用（無 Replay Buffer）	Mini-batch 多輪 epoch 更新	Replay buffer 可重複使用樣本
穩定性機制	GAE baseline 減少 variance	Clipping 限制更新幅度，或監控 KL 散度	Target network + replay buffer + OU noise
探索機制	由策略分布產生內建隨機性	同左	額外加入噪音（OU 或 Gaussian）
優勢	架構簡單、適合快速實驗	更新平穩，樣本效率高	控制精細，樣本效率極佳、能處理高維連續任務
主要限制	樣本效率低，易受高方差影響	計算量大，參數較多	初始化敏感、穩定性差、對 regime shift 抵抗力低

表 1: A2C、PPO、DDPG 三種強化學習演算法的比較

二、環境設計（胡益鳴編輯）

(一) 特徵工程：從「海量指標」到「核心訊號」

在正式訓練強化學習代理人之前，我們先面臨一個金融資料問題　特徵維度過高。若直接將所有常見技術指標餵給模型，不僅容易放大雜訊、降低收斂效率，還可能造成過擬合。為此，我們採用 Random Forest Classifier 進行重要度排序：先將每支股票下一日報酬方向（上漲 = 1、下跌 = 0）設定為標籤，再用隨機森林衡量各指標對正確分類的貢獻度。結果顯示，MACD、RSI、CCI、ADX、VR 這五項訊號最具預測力，Bollinger Bands（上下軌）與 vol_5d 亦反映出邊際資訊價值，但其重要度略低，故最終納入備用特徵池。經此篩選，狀態維度由十餘項指標縮減至重點五項，每檔股票保留五個關鍵特徵，兼顧訊號純度與運算成本。

(二) 標籤設計：動能視角的二元分類

與傳統回歸式預測「報酬率大小」不同，我們決定採用二元方向性標籤　即判斷隔日收盤價相對今日是否上漲。理由有二：其一，方向標籤符合短線交易者「做對方向即可獲利」的決策邏輯；其二，二元分類能減輕報酬率分布厚尾問題，讓模型專注萃取趨勢訊號，而非被極端值牽制。如此設計，也使隨機森林的重要度衡量更直觀，因為每一分裂節點皆以「能否提高分辨漲跌」為目標。

(三) OpenAI Gym 多資產環境：模擬真實金融市場

為了讓代理人在貼近實務的情境下學習，我們自建符合 OpenAI Gym 規格的多資產交易環境，囊括三大模組：

- 狀態空間（共 211 維）：
  - 現金餘額 1 維　反映可用流動性；
  - 30 檔股票調整收盤價 30 維；
  - 30 檔持股張數 30 維　掌握庫存；
  - 技術指標 150 維　5 個核心指標 $\times$ 30 檔股票。
- 動作空間：對每檔股票輸出介於 $[-1, +1]$ 的連續值，對應「賣出／買入比例」。真實下單前再乘以最大單次 100 股，並檢查現金與庫存約束。
- 獎勵函數：以「帳戶淨值（日終現金 + 持股市值）」的相對變動量為一步獎勵，再乘 10^{-4} 進行尺度化，避免數值爆炸。

此設計讓代理人同時面對資金分配、倉位調節與手續費三種現實限制，有效連接理論與實務落差。

(四) 環境互動邏輯：賣先於買、交易成本

每一個時間步驟，環境依序執行：

1. 接收動作向量；
2. 先賣後買　確保不會因資金不足導致未結清部位；
3. 根據當日收盤價計算實際成交金額與 0.1% 手續費，及時扣減現金；
4. 更新持股與現金後，計算 reward；
5. 回傳下一狀態與 reward，持續至資料尾端。

此流程不僅模擬撮合機制，也將交易成本內生於代理人的最適化目標，迫使策略在風險與收益之間作權衡。

(五) 策略流程：Rolling Window $\times$ Ensemble 動態選模

真實市場結構會隨時間演變。若模型長期鎖死於舊資料，性能勢必遞減。為此，我們採取 63 日（約一季）為單位的滾動式視窗，並在每一視窗內分三段：

- Train：自 2009/01/01 起至當期 Validation 前一日，分別訓練 A2C、PPO、DDPG；
- Validation：接續 63 日，根據 Sharpe Ratio 評估三種模型；

- Trade：再接續 63 日，僅用 Validation 勝出的模型下單。

期末整個視窗向前移動 63 日，重複「再訓練—再驗證—再交易」流程。如此 Ensemble 選模可在多空交錯、市場風格輪替時自動切換最佳代理人，降低單一演算法失靈風險。

(六) 模型互補性：A2C、PPO、DDPG 適合市場不同

- A2C 架構簡潔、收斂快速，適合趨勢清晰或波動有限的市場；
- PPO 透過 clipped objective 抑制過度更新，樣本效率佳、穩定度高；
- DDPG 支援精細連續動作、搭配 Replay Buffer 與 Target Network，可在橫盤時進行微調倉位。

Validation 段用 Sharpe Ratio 統一度量風險調整後績效，使得最終入選者同時兼顧收益與波動控制。

(七) 風險控制：Turbulence Index 與動態閾值

雖 RL 代理人能自學避險，但極端行情（如 COVID-19 暴跌）往往超出歷史分布，仍需額外的風險控制。我們引用 Mahalanobis Distance 為基礎的 **Turbulence Index**：

$$\text{Turbulence}_t = (p_t - \mu)^\top \Sigma^{-1} (p_t - \mu) \tag{3}$$

其中 p_t 為當日價格向量，μ、Σ 分別為 252 日均值與共變異數。數值愈高，代表當前價格脫離平常軌跡愈遠，即市場愈「動盪」。為因應結構性變化，門檻非固定，而採四步動態機制：

1. **建立長期 baseline**　先取 2009/01–2015/10 資料的 90 百分位作警戒線；
2. **每季監控**　若近 63 日平均值高於 baseline，即判定高風險；
3. **啟動 robust safety line**　高風險期改用固定閾值 96；
4. **執行端行為**　當日指標超閾即清倉並禁買，否則正常交易。

(八) 策略整合效益

整體而言，特徵篩選、滾動視窗、Ensemble 選模與 Turbulence 風控形成「訊號—決策—控風」完整閉環：

- **適應性**　訓練資料與風控 baseline 皆隨時間滾動，策略能即時吸收最新市況；
- **互補性**　三模型於 Validation 互相競爭，確保不同市場階段各有對應解法；
- **穩健性**　Turbulence 門檻在極端行情自動「踩煞車」，將大幅下跌轉化為小回撤。

肆、實證結果（李姍琪編輯）

以下表格為本次研究的實驗資訊及參數設定：

(一) 策略訓練與回測設定

項目	設定值
初始資產	$1,000,000
每檔最大交易量	100 股
Rebalance 週期	每 63 天（約每季）
交易成本	每次交易 0.1%
模型訓練結構	滾動式訓練－驗證－交易（Rolling Window）
Training 起點	2009/01/01
實際交易區間	2015/12/31–2024/10/04
訓練時長	5.84 小時

表 2: 策略訓練與回測設定參數

(二) 強化學習演算法設定

模型	Learning Rate	Episode 數
A2C	1e-4	50
PPO	1e-4	50
DDPG	Actor: 1e-3, Critic: 1e-3	50

表 3: 三種強化學習模型之學習參數設定

(三) 績效呈現（2016–2020）

為直觀呈現本研究所設計之 Ensemble 策略在歷史期間內的績效表現，以下圖表分別顯示此研究策略與道瓊指數（DJIA）於 2016 至 2020 年間的累積報酬表現以及 126 日滾動波動率變化。

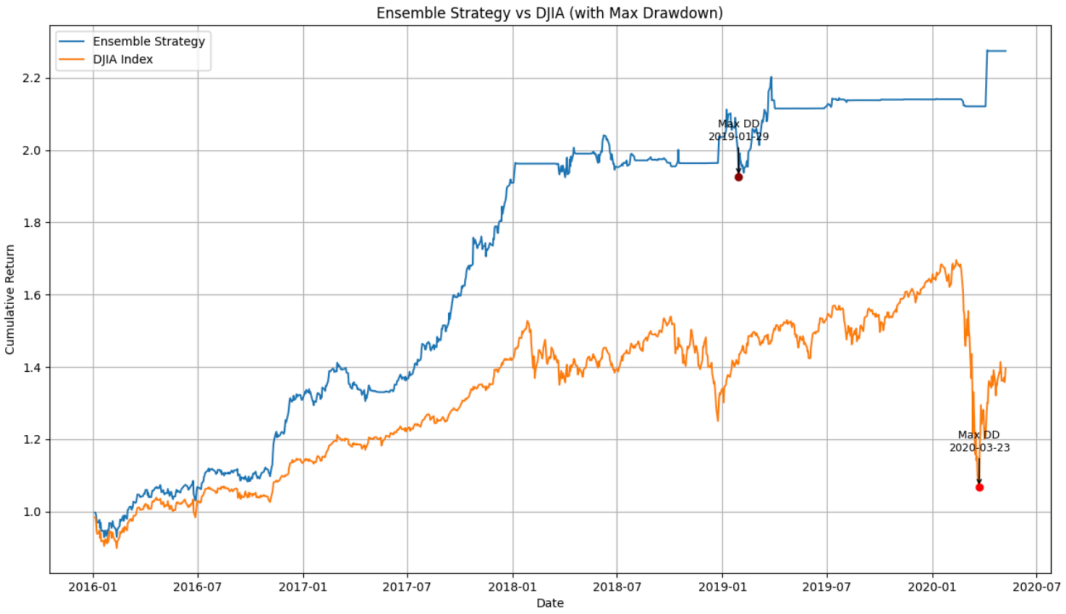

圖 2: Cumulative Return 績效呈現（2016–2020）Ensemble 策略 vs. 道瓊指數

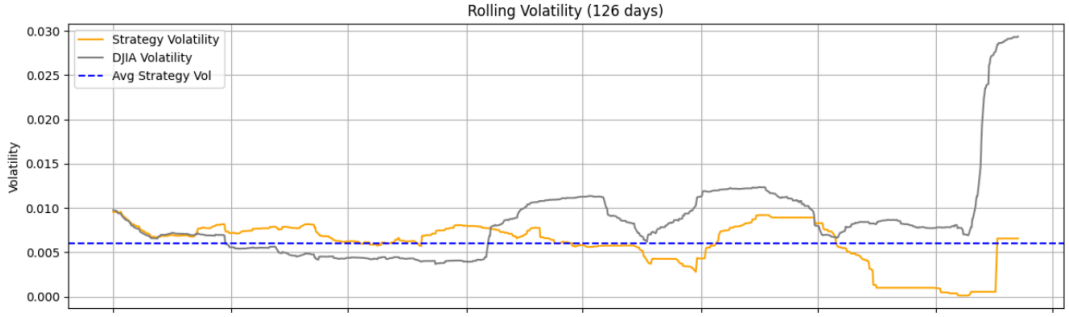

圖 3: 126 日 Rolling Volatility 績效呈現（2016–2020）Ensemble 策略 vs. 道瓊指數

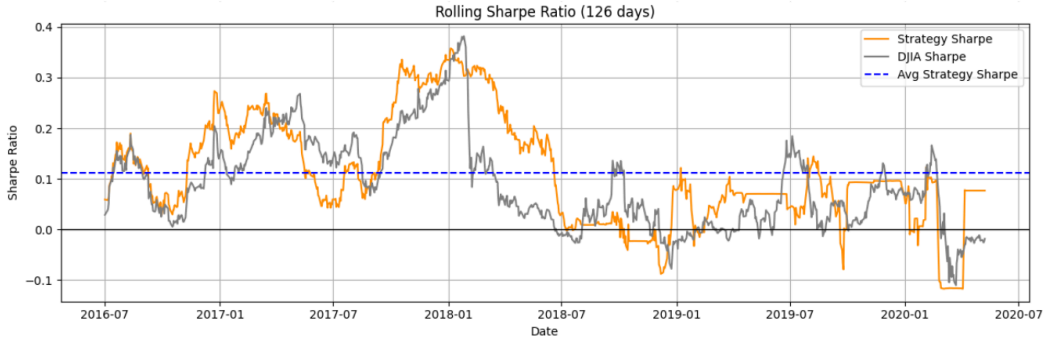

圖 4: 126 日 Sharpe Ratio 績效呈現（2016–2020）Ensemble 策略 vs. 道瓊指數

根據回測結果顯示，本研究提出的 Ensemble 策略在 2016 至 2020 年間整體表現優於道瓊指數（DJIA），自 2017 年起報酬持續上升，並於 2020 年市場劇烈波動期間仍維持穩定，展現良好抗震性與回撤控制力。

本研究策略的波動度亦長期低於道瓊指數，Sharpe Ratio 多數時間亦高於 DJIA，顯示具備穩定獲利與高風險調整後效率。

綜合五項關鍵指標之表現如下表：

指標項目	本研究 Ensemble 策略	原論文策略	DJIA 指數
累積報酬率	127.4141%	70.4%	38.6%
年化報酬率	20.79%	13%	7.8%
Sharpe Ratio	1.7837	1.3	0.47
年化波動率	10.9374%	9.7%	20.1%
最大回撤	-8.7839%	-9.7%	-37.1%

表 4: 五項關鍵績效指標之比較（2016–2020）

(四) 績效呈現（2016–2024）

為全面評估策略的長期穩定性與適應能力，本研究進一步將回測期間延長至 2024 年進行測試。以下圖表與指標呈現 Ensemble 策略在不同市場情境下的整體表現與風險控管效果。

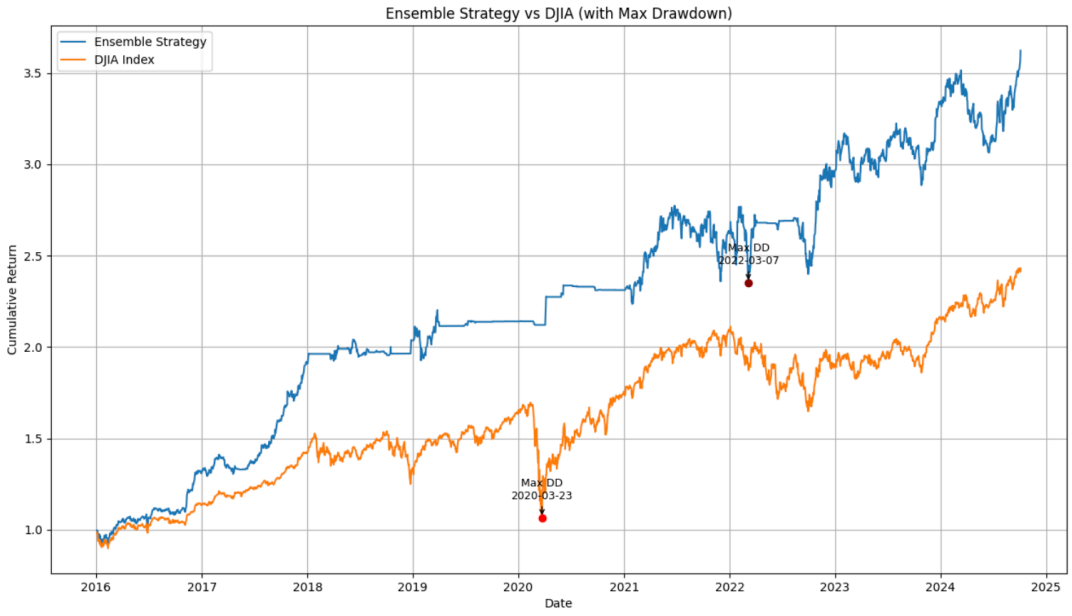

圖 5: Cumulative Return 績效呈現（2016–2024）Ensemble 策略 vs. 道瓊指數

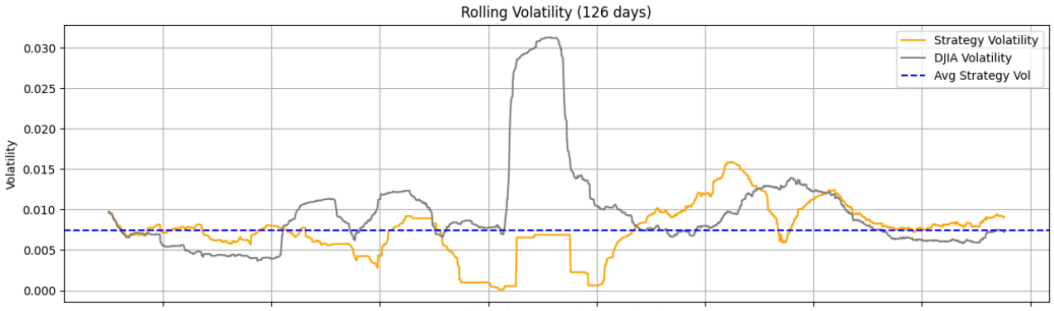

圖 6: 126 日 Rolling Volatility 績效呈現（2016–2024）Ensemble 策略 vs. 道瓊指數

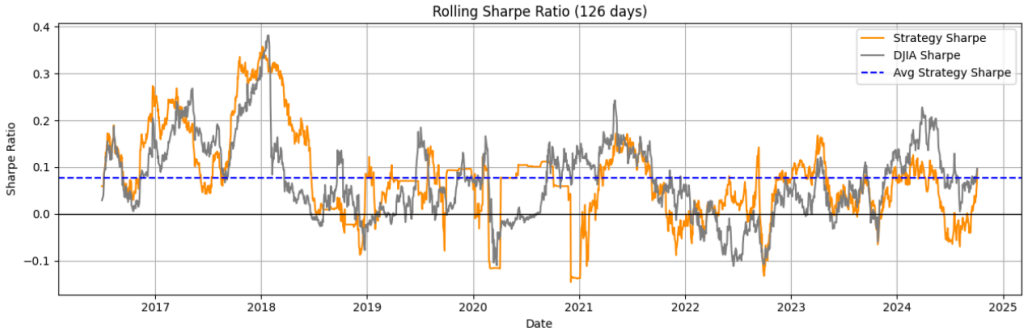

圖 7: 126 日 Sharpe Ratio 績效呈現（2016–2024）Ensemble 策略 vs. 道瓊指數

2016 至 2024 年回測結果顯示，本研究的 Ensemble 策略整體表現持續優於道瓊工業指數（DJIA）。即使在 2020 年疫情與 2022 年升息等市場動盪期間，仍能有效控制回撤並快速恢復，展現良好抗震性。最終累積報酬達 3.6 倍，遠高於 DJIA 的 2 倍。

在 126 日滾動風險指標方面，Ensemble 策略的波動度多數時間低於 DJIA，特別是在疫情與升息等高風險時期亦保持穩定，顯示良好風控能力；Sharpe Ratio 長期高於指數，說明策略能在相同風險下創造更佳報酬。

綜合五項關鍵指標之表現如下表：

指標項目	**本研究 Ensemble 策略**	**DJIA 指數**
累積報酬率	262.1638%	143.0570%
年化報酬率	15.8444%	10.6830%
Sharpe Ratio	1.1931	0.66
年化波動率	13.0469%	17.86%
最大回撤	-15.1474%	-37.0862%

表 5: 五項關鍵績效指標之比較（2016–2024）

伍、結論（李姍琪編輯）

(一) 與原研究比較

本研究參考 Yang et al. (2020) 提出的深度強化學習股票交易策略，實作並優化三種演算法（A2C、PPO、DDPG）所組成的 Ensemble 策略，搭配特徵工程、風險控制與滾動視窗等模組，設計同時具適應性及穩定性的強化學習交易系統。相較於原論文，本研究進行了以下幾項優化：

- **特徵選擇**：運用 Random Forest（隨機森林）從多種技術指標中篩選出對預測最具影響力的前五名（MACD、RSI、CCI、ADX、VR），以提升模型有效性與收斂速度；
- **風控機制**：由原論文的固定式 insample threshold 改為動態滾動門檻設計，使模型能根據當前市場波動更加彈性的調整風險暴露，有效因應如 COVID-19 疫情或升息等極端情境；
- **策略架構**：採用 rolling window 結合 Ensemble 選模機制，使模型具備持續學習與切換 agent 的能力，降低單一模型失效的風險。

整體而言，優化後的策略在各項報酬與風險衡量指標上均優於原始論文及市場基準（DJIA），證明本研究架構具備實務應用潛力。

(二) 本研究未來改善方向

本研究已初步建構出具備穩健性與彈性的強化學習股票交易策略，然而在策略細節與架構層面，未來的優化方向將聚焦於以下兩大面向：

- **特徵工程**：未來可持續擴充技術指標與市場變數的納入範圍，並考慮引入如新聞情緒分析、宏觀經濟數據、資金流向等非結構化資訊，以強化模型對市場變化的前瞻性判斷；
- **Reward Function**：期待能向更具市場行為解釋力的方向發展，減少對硬性風控條件（如閾值設定）的依賴。透過讓 agent 在學習過程中自行辨識市場風險並作出相應調整，建立具備內建風控能力的自主決策模型。

陸、參考文獻

Yang, H., Liu, X.-Y., Zhong, S., & Walid, A. (2020). Deep Reinforcement Learning for Automated Stock Trading: An Ensemble Strategy. *Proceedings of the First ACM International Conference on AI in Finance (ICAIF' 20), October 15–16, 2020, Manhattan, NY, USA*. ACM, New York, NY, USA.

最後由黃得晉排版整理轉成