

國立陽明交通大學
資訊管理與財務金融學系財務金融碩士班
碩士論文

Graduate Program of Finance
National Yang Ming Chiao Tung University
Master Thesis

基於圖神經網絡的股票收益預測與選股策略—融合市場結構的多因子模型
Stock Return Prediction and Selection via Graph Neural Networks: Integrating Market Structure into a Multi-Factor Model

研 究 生：羅頤 (Lo, Yi)
指導教授：黃宜侯 (Huang, Alex YiHou)

中華民國 一一五年六月
June 2026

基於圖神經網絡的股票收益預測與選股策略—融合市場結構的多因子模型

Stock Return Prediction and Selection via Graph Neural Networks:
Integrating Market Structure into a Multi-Factor Model

研究 生： 羅頤

Student: Yi Lo

指導教授： 黃宜侯 博士

Advisor: Dr. Alex YiHou Huang



June 2026
Taiwan, Republic of China

中華民國 一一五年六月

誌 謝

謝天謝地，感謝國立陽明交通大學提供良好的研究環境與資源，使我能順利完成學業。



基於圖神經網絡的股票收益預測與選股策略—融合市場結構的多因子模型

學生：羅頤

指導教授：黃宜侯 博士

國立陽明交通大學
資訊管理與財務金融學系財務金融碩士班

摘要

中文摘要就從這邊開始寫。

本研究採用圖注意力網絡 (GAT) 和動態多因子模型 (DMFM)，利用產業圖和全市場圖的結構特性進行個股收益率預測，並通過動態投資組合回測評估模型的實際投資績效。

關鍵字：圖注意力網絡、股票收益率預測、量化投資策略

Stock Return Prediction and Selection via Graph Neural Networks: Integrating Market Structure into a Multi-Factor Model

Student : Yi Lo

Advisor: Dr. Alex YiHou Huang

Graduate Program of Finance
National Yang Ming Chiao Tung University

Abstract

Write your English abstract here. This research employs Graph Attention Networks (GAT) and Dynamic Multi-Factor Models (DMFM) to predict individual stock returns using structural characteristics of industry graphs and market-wide graphs, and evaluates the practical investment performance through dynamic portfolio backtesting.

Keywords: Graph Attention Network, Stock Return Prediction, Quantitative Investment Strategy.

目錄

中文摘要.....	i
英文摘要.....	ii
目錄.....	iii
圖目錄.....	vi
表目錄.....	vii
第一章 緒論.....	1
1.1 研究背景.....	1
1.2 研究問題.....	2
1.3 研究方法與貢獻.....	3
1.4 論文架構.....	4
第二章 文獻回顧.....	5
2.1 技術指標如何轉化為深度學習可用之特徵表示.....	5
2.2 LSTM 等時間序列模型在金融預測中的典型做法與延伸.....	6
2.3 圖神經網路的基本概念與代表性架構.....	8
2.4 GAT 及其變形在股票預測／排序任務上的應用與限制.....	8

第三章 實驗設計.....	10
3.1 資料來源與預處理.....	10
3.1.1 資料來源.....	10
3.1.2 TEJ 資料欄位與研究變數表格	11
3.1.3 資料預處理.....	11
3.1.4 資料集劃分.....	12
3.2 特徵工程與圖結構.....	13
3.2.1 特徵選取與分類.....	13
3.2.2 圖結構設計與建構流程.....	14
3.3 模型設計.....	14
3.3.1 基準模型: GATRegressor.....	14
3.3.2 DMFM 模型架構 (Wei et al. 2022)	15
3.4 損失函數.....	16
3.4.1 總損失函數.....	16
3.5 評估指標.....	16
3.5.1 預測品質指標.....	16
3.5.2 投資組合績效指標.....	17
3.5.3 基準比較與流程摘要.....	17
第四章 實驗結果.....	18
4.1 模型對比.....	18
4.2 投組績效驗證.....	18

4.3 特徵重要性排名.....	18
第五章 結論.....	19
5.1 主要發現.....	19
5.2 與既有研究的對比.....	19
5.3 模型限制.....	19
5.4 實踐應用與未來方向.....	19
參考文獻.....	20
Appendix 附錄.....	22



圖目錄

圖 1 LSTM 單元的資訊流 (Forget / Update / Output)。其中 x_t 為當期輸入，

h_{t-1} 為前一期隱狀態 (hidden state), c_{t-1} 為前一期記憶狀態 (cell state)。

6



表目錄

表 1 TEJ 匯出之原始欄位（資料來源欄位）	11
表 2 模型輸入特徵摘要（完整 54 項清單見附錄表 3）	13
表 3 本研究實際使用之模型輸入特徵（feature_cols，共 54 項；全部由 TEJ 原始欄位衍生）	22



第一章、緒論

1.1 研究背景

股票市場的價格形成同時受到公司基本面、產業景氣、整體市場風險偏好與短期交易行為等因素影響，使得報酬序列呈現高雜訊、非線性與時變（time-varying）特性。在量化投資（quantitative investment）實務中，常以可計算的市場訊號作為決策依據，其中技術指標（Technical Indicators, TIs）透過歷史價格與成交量序列，將趨勢、動能與波動等資訊轉換為量化特徵，長期被廣泛用於特徵工程（feature engineering）與交易策略建構。

近年深度學習（deep learning）在金融預測任務上的應用逐漸普及，主要優勢在於能以資料驅動方式學習非線性映射，並在高維特徵下自動萃取有效表示。然而，當技術指標與衍生特徵數量擴張時，特徵集合往往伴隨高度相關、冗餘與尺度不一致，導致模型訓練成本提高、收斂不穩定，並可能削弱泛化能力。換言之，研究焦點逐步由「是否使用技術指標」轉向「如何將技術指標組織為深度模型可有效吸收的表示」。

另一方面，金融市場並非由彼此獨立的資產所構成。台灣股票之間普遍存在產業層級與市場層級的共同因子影響（common factors），使個股報酬在截面（cross-section）上呈現結構性相依。傳統時間序列模型（如 LSTM）多以單一資產的歷史序列為主要訊號來源，較難直接刻畫股票之間的關係拓撲；因此近年研究開始引入圖神經網路（Graph Neural Networks, GNNs），以「節點（股票）—邊（關係）」的方式建模產業或市場連結，並透過訊息傳遞（message passing）與注意力機制（attention mechanism）聚合鄰居資訊，以提升截面預測與排序任務的表現與可解釋性。

在上述脈絡下，結合多因子表示（multi-factor representation）與圖注意力（graph attention）的模型架構，特別是能分離產業與全市場共同影響的設計，提供了一條兼顧「特徵可學習性」與「關係結構建模」的路徑。本文後續將採用深度多因子模型結合圖注意力之框架，並以台灣日頻資料進行實證檢驗。

1.2 研究問題

綜合既有研究與實務需求，本文關注的核心問題可整理如下。

第一，技術指標所形成的高維特徵集合雖能涵蓋多面向市場訊號，但亦容易引入冗餘與尺度差異，使模型難以穩定學習；因此需要一套可重複、可檢核的特徵建置與整理流程，將價格與量能序列轉換為適合深度模型輸入的特徵表示，並降低不必要的噪音干擾。

第二，股票之間存在產業共通訊號與市場共同波動，僅依賴單一股票的時間序列特徵，可能無法充分利用截面關係資訊；因此需要能顯式建模股票關係的學習架構，使模型同時吸收「個股特徵」與「關係拓撲」兩類訊號。

第三，即使引入圖結構，模型表現仍高度仰賴圖的構建方式與關係品質；若未能有效分離產業與全市場層級的共同影響，模型可能將「產業押注」或「市場曝險」誤判為選股能力。因而本文特別關注：是否能透過分層中性化（hierarchical neutralization）的方式，將產業與市場影響自特徵表示中拆解，進而提升預測訊號的穩健性。

第四，在金融預測情境下，模型品質不宜僅以單一誤差指標衡量；更關鍵的是模型在截面排序上的有效性與穩定性。因此本文採用資訊係數（Information Coefficient, IC）與其穩健性指標 ICIR（Information Coefficient Information Ratio, ICIR）作為主要評估基礎，並輔以投資組合績效指標進行整體驗證。

1.3 研究方法與貢獻

本文採用台灣股票日頻資料進行實證。資料主要來源為 TEJ 台灣經濟新報 (Taiwan Economic Journal, TEJ)，並匯出為 CSV 作為後續處理之輸入。以「交易日 × 股票」為觀測單位，保留日期、股票代碼、OHLC、成交量／成交值、市值與產業分類等欄位；技術特徵則由價格與量能序列計算而得，其中 RSI、MACD、KD 等指標透過 TA-Lib 套件產生，最終形成固定的 54 維特徵集合供模型輸入（詳見第三章的變數表格與附錄清單）。

在關係建模方面，本文依第三章之設計建構兩類圖結構：(1) 產業圖 (Industry Graph)，以同產業股票完全連結並加入自環；(2) 全市場圖 (Universe Graph)，以完全圖表示整體市場共同因子影響。圖結構以 PyTorch Geometric (PyG) 的 edge_index 形式保存，並與特徵張量共同形成可重複使用的訓練產物 (artifacts)。

模型設計以 Deep Multi-Factor Model (DMFM) 為主體，透過「特徵編碼 → 產業中性化 → 全市場中性化 → 階層式拼接 → 深度因子學習 → 因子注意力」的流程，在每個交易日的截面上輸出各股票之預測訊號；同時以較簡化的圖模型 (GATRegressor) 作為基準模型，進行可比性的對照實驗。評估上以 IC、Daily IC、ICIR 與產業中性 IC 作為核心預測指標，並以投資組合回測指標（年化報酬率、Sharpe ratio、勝率等）檢驗策略層級的有效性；資料切分採時間序列切分以避免資訊洩漏，並設計不同樣本長度（短期／中期／長期）檢驗穩健性。

基於上述方法設計，本文的主要貢獻歸納如下：

- **資料與特徵流程的可重現性：**以 TEJ 日頻資料為基礎，建立從原始欄位到 54 維技術特徵（含 TA-Lib 指標）的完整特徵建置與張量化流程，並以 artifacts 形式保存，便於重複訓練與評估。
- **關係結構的分層建模：**同時建構產業圖與全市場圖，並在模型中引入分層中性化機制，以拆解產業與市場共同影響，提升截面預測訊號的穩健性與可解釋性。

- **以排序導向指標為核心的實證驗證：**以 IC/ICIR 與產業中性 IC 作為主要評估基礎，並搭配投資組合績效指標，提供模型在台灣市場情境下的多面向驗證結果。

1.4 論文架構

本文共分為五章，各章內容如下。

第一章為緒論，說明研究動機與背景、研究問題、研究方法與預期貢獻，並概述全文架構。

第二章為文獻回顧，依序整理技術指標在深度學習中的特徵表示、時間序列模型（如 LSTM）於金融預測的典型做法、圖神經網路的基本概念與代表性架構，以及圖注意力模型（GAT）在股票預測／排序任務上的應用與限制，作為後續方法設計之理論與實證依據。

第三章為實驗設計，說明資料來源與預處理流程、特徵工程與圖結構建構方式、模型架構（基準模型 GATRegressor 與 DMFM）、損失函數設定，以及評估指標與實驗切分規則。

第四章為實驗結果，呈現不同模型之預測表現比較、IC / ICIR 與產業中性 IC 之結果分析，並以投資組合績效驗證模型訊號在策略層面的有效性與穩健性（含不同樣本長度設定之比較）。

第五章為結論，總結本文之主要發現，討論方法限制與可能的改進方向，並提出後續在實務應用與研究延伸上的建議。

第二章、文獻回顧

本章依序回顧：（1）技術指標如何轉化為深度學習可用之特徵表示，（2）LSTM 等時間序列模型在金融預測中的典型做法與延伸，（3）圖神經網路的基本概念與代表性架構，（4）GAT 及其變形在股票預測／排序任務上的應用與限制，並據此銜接本研究後續以多因子與圖注意力為核心之模型設計與實驗流程。

2.1 技術指標如何轉化為深度學習可用之特徵表示

技術指標（Technical Indicators, TIs）以歷史價格與成交量為基礎，將市場的趨勢、動能與波動等訊號轉換為可計算的量化特徵，長期被用於股票預測任務的特徵工程。當指標種類擴張時，特徵集合往往同時伴隨高度相關、冗餘與尺度不一致，使模型訓練成本上升，並可能影響泛化能力與穩健性。

為回應上述挑戰，研究焦點逐步由「是否使用技術指標」轉向「如何將技術指標組織為深度學習可有效吸收的特徵表示」。Agrawal 等人以技術指標作為深度學習模型的輸入，並在股價預測任務中強調多指標組合能提供更完整的市場訊號。^[1] Agrawal 等人亦提出以最佳化的深度學習架構搭配技術指標的預測流程，指出在輸入維度增加時，指標集合的挑選與組織方式將直接影響模型是否能有效吸收訊號並降低冗餘干擾。^[2] 此外，Li 與 Bastos 以系統性綜述方式整理「技術分析十深度學習」在股市預測的主要設計與趨勢，並指出「如何識別最優指標集」仍是領域內的重要議題。^[3]

總結而言，技術指標在深度學習框架中的角色可被視為「可學習特徵空間的原料」：研究重點不僅在於指標本身，而在於如何以更合適的表示方式降低冗餘並保留有效訊

號，進而提升後續預測任務的可學習性與穩定性。

2.2 LSTM 等時間序列模型在金融預測中的典型做法與延伸

時間序列模型在金融預測中主要利用價格與特徵序列的時間依賴性，學習由過去資訊推估未來表現。長短期記憶網路（Long Short-Term Memory, LSTM）透過閘門（gates）機制控制資訊流動，能緩解長序列學習的梯度消失（vanishing gradients）問題，因此成為金融時間序列建模的代表性方法之一。

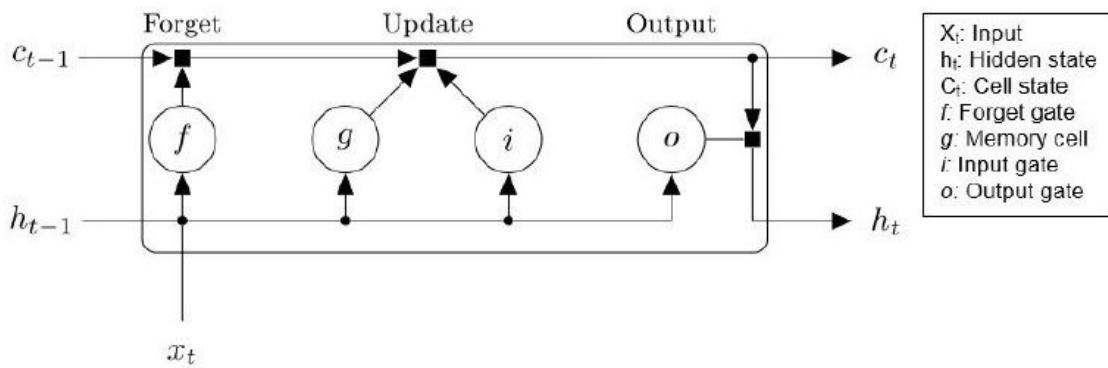


圖 1: LSTM 單元的資訊流 (Forget / Update / Output)。其中 x_t 為當期輸入， h_{t-1} 為前一期隱狀態 (hidden state)， c_{t-1} 為前一期記憶狀態 (cell state)。

如圖 1 所示，LSTM 會沿著記憶狀態 (cell state) 主幹 $c_{t-1} \rightarrow c_t$ 保留長期資訊，並透過三個閘門控制「遺忘 (Forget)」、「更新 (Update)」與「輸出 (Output)」：(1) 遺忘閘門 f_t 決定保留多少過去記憶 c_{t-1} ；(2) 更新步驟由輸入閘門 i_t 與候選記憶 g_t (圖中 memory cell) 共同決定要寫入多少新資訊；(3) 輸出閘門 o_t 決定由當期記憶 c_t 產生多少輸出到

隱狀態 h_t 。對應的計算可寫為：

$$\begin{aligned} f_t &= \sigma(W_f[h_{t-1}, x_t] + b_f), \\ i_t &= \sigma(W_i[h_{t-1}, x_t] + b_i), \\ g_t &= \tanh(W_g[h_{t-1}, x_t] + b_g), \\ c_t &= f_t \odot c_{t-1} + i_t \odot g_t, \\ o_t &= \sigma(W_o[h_{t-1}, x_t] + b_o), \\ h_t &= o_t \odot \tanh(c_t), \end{aligned} \tag{2.1}$$

其中 $\sigma(\cdot)$ 為 sigmoid 函數、 $\tanh(\cdot)$ 為雙曲正切函數， \odot 表示逐元素相乘 (element-wise multiplication)。圖中的方形節點可視為逐元素乘法與加總的運算節點，對應到 $f_t \odot c_{t-1}$ 與 $i_t \odot g_t$ 兩條路徑匯入 c_t 的更新。

在實務與研究中，LSTM 的延伸方向常包含多目標預測、結合注意力機制 (attention mechanism) 與去噪／重加權策略，以及高頻資料下的深層序列建模。Zaheer 等人設計多參數預測架構以同時預測不同價格維度，並比較不同深度模型配置在特定資料規模下的表現差異。^[4] Qiu 等人將注意力機制引入 LSTM 架構，以在時間維度上動態調整特徵權重，強化模型在金融資料高雜訊情境下的預測能力。^[5] 在高頻（例如 5 分 K）資料情境下，鄭邦廷以疊層式 LSTM (Stacked LSTM) 捕捉更複雜的非線性時間依賴，用於買賣點預測。^[6] 此外，廖俊翔透過自相關分析與特徵篩選納入跨市場外部特徵，並展示外部訊號在一定程度上能降低預測誤差。^[7]

整體而言，時間序列方法能有效吸收單一資產（或單一特徵集合）的歷史資訊，但其多數設計仍以「序列本身」為主要訊號來源，較難直接刻畫股票之間的結構性關係與共通影響；因此後續研究開始引入圖結構以建模市場關係。

2.3 圖神經網路的基本概念與代表性架構

圖神經網路 (Graph Neural Networks, GNNs) 以圖作為基本資料結構，透過節點 (node) 與邊 (edge) 描述實體及其關係，並以訊息傳遞 (message passing) 機制聚合鄰居資訊，學得具備拓撲語意的節點表示。Wu 等人對 GNN 架構進行系統性整理，並區分頻域 (spectral-based) 與空域 (spatial-based) 圖卷積之主要差異，同時歸納多類代表性 GNN 架構族群。[8]

在金融市場中，股票之間存在產業層級與市場層級的共同因子影響，亦可能呈現共動性、傳染效應與結構性相依關係。以圖結構建模股票關係，可使模型同時利用「個股特徵」與「關係拓撲」來提升截面預測或排序任務的表現，亦為後續在模型中引入中性化或分離共通影響提供方法基礎。

2.4 GAT 及其變形在股票預測／排序任務上的應用與限制

圖注意力網路 (Graph Attention Network, GAT) 在鄰居聚合框架中引入注意力機制，使模型可對不同鄰居節點賦予不同權重，提升在異質或噪音關係下的表徵能力與可解釋性。以注意力係數為例，其形式可寫為：

$$\alpha_{ij} = \text{softmax}_j \left(\text{LeakyReLU} \left(a^\top [Wh_i \parallel Wh_j] \right) \right), \quad (2.2)$$

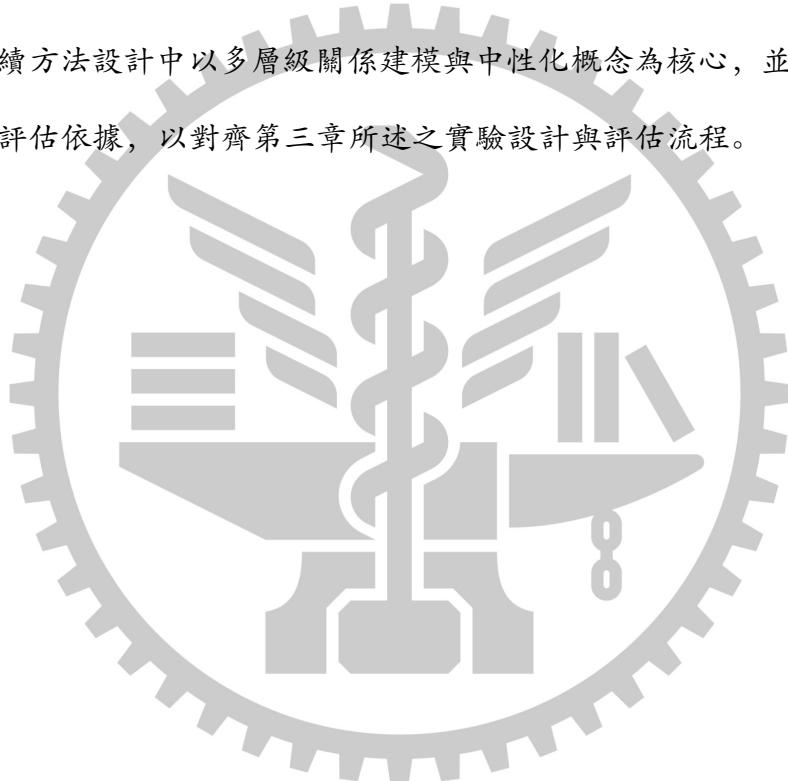
其中 h_i 為節點表示， α_{ij} 代表節點 i 對鄰居 j 的注意力權重。

在股票預測任務中，GAT 類方法的效能高度依賴圖結構建構方式與關係訊號品質。Huang 等人提出多層級圖注意力模型 (ML-GAT)，以分層注意力機制分別處理節點狀態與關係類型，使模型能在多來源關係下更細緻地擷取有效訊號。[9] Song 等人以股票價格與股票關係資訊進行圖聚合式排序預測，並討論關係資訊設計對截面排序表現的影

響。[10] 當市場關係稀疏或不完整時，Cheng 等人提出多特徵圖注意力網路，透過整合多面向特徵訊號來提升股票預測能力。[11]

然而，GAT 類方法仍可能面臨下列限制：其一，注意力權重在噪音邊存在時可能分散，增加過擬合風險；其二，圖建構規則、關係來源與連邊密度高度敏感，使模型可轉移性與穩健性需要透過更嚴謹的實證評估驗證；其三，靜態圖難以全面反映關係隨時間變動的特性，可能降低模型對市場結構變化的敏感度。

基於上述研究脈絡，Wei 等人提出以深度多因子模型結合圖注意力的框架，並透過分離產業與全市場層級影響來提升因子信號的可用性與穩健性。[12] 本研究將沿用此一方向，在後續方法設計中以多層級關係建模與中性化概念為核心，並以 IC／ICIR 等指標作為實證評估依據，以對齊第三章所述之實驗設計與評估流程。



第三章、實驗設計

本章說明整體實驗流程，涵蓋資料來源與預處理、特徵工程與圖結構建構、模型架構與訓練設定、損失函數與評估指標。實驗目標在於檢驗 Deep Multi-Factor Model (DMFM) 於台灣股票市場之預測表現，並與基準模型進行對照。

3.1 資料來源與預處理

3.1.1 資料來源

資料取自 **TEJ 台灣經濟新報 (Taiwan Economic Journal, TEJ)** 之台灣股票市場日頻資料，並以 CSV 形式匯出以供後續處理與模型訓練使用。資料以「交易日 × 股票」為觀測單位，至少包含日期 (Date)、股票代碼 (Stock ID) 與收盤價 (Close) 等必要欄位；同時保留開高低收 (OHLC)、成交量／成交值、市值、流通股數、估值指標（如 PB、PS）與產業分類等資訊，以支援特徵建置與圖結構構建。樣本期間由參數 start_date 與 end_date 控制，並以台股交易日序列作為時間索引基準。

除 TEJ 可直接取得之原始欄位外，模型輸入之技術特徵由價格與成交量等基礎序列計算而得；其中 RSI、MACD、KD 等指標主要透過 TA-Lib 套件生成，再彙整為固定的特徵集合供模型輸入。為避免於文字中重複列舉欄位與變數定義，TEJ 原始欄位彙整於表 1，最終模型輸入之 54 個特徵摘要見表 2（完整清單見附錄表 3）。

3.1.2 TEJ 資料欄位與研究變數表格

所有輸入資料均由 TEJ 匯出為日頻 CSV。匯出檔案包含：(1) 市場交易與公司屬性之原始欄位（表 1），以及 (2) 由程式以 TEJ 原始欄位衍生之技術特徵集合 feature_cols ($n=54$)（表 2）。

表 1: TEJ 匯出之原始欄位（資料來源欄位）

TEJ 欄位名稱	用途	備註（單位/說明）
年月日	時間索引	交易日 (Date)
開盤價 (元)	交易價格	Open (元)
最高價 (元)	交易價格	High (元)
最低價 (元)	交易價格	Low (元)
收盤價 (元)	交易價格	Close (元)
成交量 (千股)	交易量能	Volume (千股)
成交值 (千元)	交易金額	Amount (千元)
本益比-TEJ	估值資訊	Price-to-Earnings (TEJ 口徑)
股價淨值比-TSE	估值資訊	Price-to-Book (TSE 口徑)
證券代碼_純代碼	個股識別	Stock ID
證券名稱	個股識別	公司名稱
TEJ 產業_名稱	產業資訊	產業分類 (TEJ 分類；用於建構產業圖)

3.1.3 資料預處理

資料預處理主要涵蓋資料格式統一、特徵與標籤計算、以及張量化整理，並以一致規則處理缺失值以降低訓練偏誤。整體流程如演算法 1 所示。

Algorithm 1 資料預處理流程

- 1: 讀取 TEJ 匯出之日頻資料（至少包含 Date、Stock ID、Close；其餘欄位供特徵與圖結構使用）
 - 2: 依 Stock ID 與 Date 排序；進行日期格式與數值型態轉換，並做基本缺漏/異常檢核
 - 3: 依價格與量能序列計算技術特徵（含 TA-Lib 指標），形成 feature_cols
 - 4: 計算標籤: $y_t = \frac{P_{t+k} - P_t}{P_t}$ # k 日未來報酬率 (forward return)
 - 5: 由產業欄位建構產業圖 E_{ind} ；由股票清單建構全市場圖 E_{uni}
 - 6: 建立特徵張量 $F^t \in \mathbb{R}^{T \times N \times F}$ 與標籤張量 $y^t \in \mathbb{R}^{T \times N}$
 - 7: 特徵 NaN/Inf 以 0 取代；標籤 NaN 保留並於訓練/評估以 mask 排除
 - 8: 特徵正規化由模型端 Batch Normalization (BatchNorm) 處理
-

缺失值處理 (Missing values): 不對缺失值進行插補，以避免引入主觀假設。特徵層

面將 NaN/Inf 以 0 取代以維持張量完整性；標籤層面保留 NaN，並在訓練與評估時以有效樣本遮罩（mask）排除，使模型僅在可觀測標籤上學習。

特徵標準化 (Normalization)：資料前處理階段不額外做截面標準化，特徵保留原始尺度；模型端於輸入層使用 **Batch Normalization (BatchNorm)** 進行正規化：

$$x_{normalized} = \frac{x - \mu_{batch}}{\sqrt{\sigma_{batch}^2 + \epsilon}} \quad (3.1)$$

其中 μ_{batch} 與 σ_{batch}^2 為 mini-batch 統計量， ϵ 為數值穩定項（預設 10^{-5} ）。

標籤定義 (Label definition)：標籤定義為未來 k 日報酬率（forward- k return）：

$$y_t = \frac{P_{t+k} - P_t}{P_t} \quad (3.2)$$

其中 P_t 為時間 t 收盤價， k 為預測視窗（預設 $k = 5$ ）。

3.1.4 資料集劃分

採時間序列切分以避免資訊洩漏（look-ahead bias）。資料依時間順序切分為訓練集與測試集，比例為 80%:20%（前 80% 為訓練、後 20% 為測試）。另為檢驗不同樣本長度之下之穩健性，設計短期／中期／長期三種資料長度，並皆採用相同的 80%/20% 時間切分：

- **短期資料**：2019-09-16 ~ 2020-12-31（共 319 個交易日；訓練 255 日、測試 64 日）
- **中期資料**：2019-09-16 ~ 2022-12-31（共 809 個交易日；訓練 647 日、測試 162 日）
- **長期資料**：2019-09-16 ~ 2025-09-12（共 1460 個交易日；訓練 1168 日、測試 292 日）

3.2 特徵工程與圖結構

3.2.1 特徵選取與分類

模型輸入特徵涵蓋報酬、趨勢、波動、量能、流動性與技術指標等訊號來源，並以多視窗（rolling window）構建不同時間尺度之市場資訊。各特徵之**實際使用變數名稱、定義、視窗長度與對應 TEJ 原始欄位**已於附錄表 3 完整列示。除 TEJ 匯出欄位外，技術指標部分由 TA-Lib 依價格與量能序列計算，不另引入外部第三方因子資料。

表 2: 模型輸入特徵摘要（完整 54 項清單見附錄表 3）

類別	代表特徵	定義（由 TEJ 欄位計算）	視窗
報酬	ret	$P_t/P_{t-k} - 1$ (TEJ 收盤價)	1/5/20
反轉	rev	$-ret_k$ (TEJ 收盤價)	1/5/10
趨勢	px_over_sma	$P_t/SMA_k(P)_t - 1$ (TEJ 收盤價)	5/20/60
動能差	mom_diff	$ret_{k_1} - ret_{k_2}$ (例如 10 – 1, 20 – 1; – TEJ 收盤價)	–
波動	std_ret	$Std(r_{t-k+1}, \dots, r_t)$ (TEJ 收盤價)	5/20/60
分配	skew, kurt	20 日報酬偏度/峰度 (TEJ 收盤價)	20
量能	vol_over_ma	$V_t/SMA_k(V)_t - 1$ (TEJ 成交量)	5/20
量能	up_vol, down_vol	上/下漲日平均量 (TEJ 成交量；以日報酬正負分組)	20
流動性	amihud	$\frac{1}{k} \sum \frac{ r }{Amt+\epsilon}$ (TEJ 成交值)	5/20
技術指標	rsi, macd, stoch_k	RSI / MACD / KD (以 TEJ 序列由 TA-Lib 計算)	14/20
風險	atr, mdd	ATR (TEJ 高低收) / 最大回撤 (TEJ 收盤價)	14/20
風險	beta, idio_vol	rolling beta / 特質波動 (以個股與市場報酬 rolling 估計)	60
區間極值	roll_max, roll_min	rolling max/min (TEJ 收盤價)	5/10/20/60
區間位置	pct_pos	$\frac{P_t - \min(P)}{\max(P) - \min(P) + \epsilon}$ (TEJ 收盤價)	5/10/20/60
位置差距	pct_to_high, pct_to_low	距離/低點比例 (以區間 max / min 與 P_t 計；TEJ 收盤價)	20
標準化	zscore_close	$\frac{P_t - \mu_k}{\sigma_k + \epsilon}$ (TEJ 收盤價)	20/60

產業分類 (Industry classification)：產業欄位採用 TEJ 提供之 TEJ 產業 _ 名稱。

該分類於樣本期間內具相對穩定性，適合用於建構產業層級圖結構；實驗中不額外進行主觀重分群。

3.2.2 圖結構設計與建構流程

產業圖（Industry Graph）：以「同產業股票完全連結」為原則，將同一產業內之股票視為相互連結，並為每檔股票加入自環（self-loop），以捕捉產業內共通訊號。

全市場圖（Universe Graph）：採完全圖（complete graph）設計，所有股票相互連結並包含自環，用以表徵跨產業之市場共同因子影響。

表示方式（Representation）：圖結構以 PyTorch Geometric (PyG) 之 edge_index 表示（形狀為 $[2, E]$ ）。實作上先依規則建立連結關係，再轉換為 edge_index 以供 GAT 層運算。

建構流程（Construction）：由輸入 CSV 讀取產業欄位並依股票代碼分組建立產業圖；若產業欄缺漏，則以單一產業處理。全市場圖依股票清單建立完全連結。圖結構與特徵量共同保存為 artifacts，以支援訓練與評估之重複使用。

3.3 模型設計

3.3.1 基準模型：GATRegressor

基準模型採用 **GATRegressor** 作為對照：以兩層 Graph Attention Network (GAT) 於產業圖上進行訊息傳遞，並以線性層輸出預測值。相較於 DMFM，GATRegressor 不包含「產業中性化 + 全市場中性化」之階層式結構，亦不含因子注意力模組，因此可作為較簡化之圖模型基準。超參數設定盡量與 DMFM 對齊（例如 hidden dimension、heads、dropout），以提升可比性。

3.3.2 DMFM 模型架構 (Wei et al. 2022)

DMFM 之核心流程可概括為：「特徵編碼 → 產業中性化 → 全市場中性化 → 階層式拼接 → 深度因子學習 → 因子注意力」。模型在每個交易日 t 以截面方式處理所有股票之特徵，並輸出每檔股票之預測因子值。

步驟 1：特徵編碼 (Feature encoder)：原始特徵 F^t 經 BatchNorm 後輸入 MLP，映射為 C^t （維度為 `hidden_dim`），作為股票之語境表示。

步驟 2：產業中性化 (Industry neutralization)：以產業圖進行 GAT 運算得到產業影響 H_I^t ，並定義產業中性特徵為 $C_I^t = C^t - H_I^t$ 。

步驟 3：全市場中性化 (Universe neutralization)：以全市場圖在 C_I^t 上進行 GAT 運算得到 H_U^t ，並定義全市場中性特徵為 $C_U^t = C_I^t - H_U^t$ 。

步驟 4：階層式特徵拼接 (Hierarchical concatenation)：拼接三層表示形成 $H^t = [C^t \parallel C_I^t \parallel C_U^t]$ ，以同時保留原始語境與兩層中性化資訊。

步驟 5：深度因子學習 (Deep factor learning)：將 H^t 輸入 decoder MLP，輸出深度因子 f^t （每檔股票一個預測值）。

步驟 6：因子注意力 (Factor attention)：模型同時學習特徵注意力權重 a^t ，形成 $\hat{f}^t = F^t \odot a^t$ ，用以估計驅動因子輸出之重要特徵，並作為監督與解釋性分析依據。

超參數配置 (Hyperparameters)：預設為 `hidden_dim=64`、`heads=2`、`dropout=0.1`、`epochs=200`、`learning rate=1e-4`、`weight decay=0.01`、 $\lambda_{attn} = 0.1$ 、 $\lambda_{IC} = 1.0$ 、`patience=30`。

3.4 損失函數

3.4.1 總損失函數

採用 DMFM 對應之損失函數：

$$L = \lambda_{attn} \cdot \|f - \hat{f}\| + \lambda_{IC} \cdot (1 - IC) - \lambda_b \cdot b \quad (3.3)$$

其中 $\|f - \hat{f}\|$ 為注意力估計誤差， IC 為資訊係數（Information Coefficient, IC）， b 為截面回歸得到的因子收益項。實作中 λ_b 固定為 0.01，以降低因子收益項之不穩定性。

IC (Information Coefficient): 以 Pearson correlation 計算，基於每個交易日之截面預測與真實 forward- k 日報酬，衡量預測排序有效性。

優化器 (Optimizer): 使用 AdamW，並加入 weight decay 以抑制過度擬合。

3.5 評估指標

3.5.1 預測品質指標

IC / Daily IC / ICIR: IC 為整體樣本之 Pearson correlation；Daily IC 為逐交易日 IC 之平均；ICIR (Information Coefficient Information Ratio, ICIR) 定義為 Daily IC 平均除以其標準差，用以衡量穩健性。

產業中性 IC (Industry-neutral IC): 為排除產業共同波動之影響，先在每交易日、每產業內分別對預測與真實報酬去均值 (industry de-meaning)，再對殘差計算 Pearson

correlation。令 $\hat{y}_{i,t}$ 與 $y_{i,t}$ 分別為股票 i 在日 t 的預測與真實報酬， $g(i)$ 表示產業別，則

$$\tilde{\hat{y}}_{i,t} = \hat{y}_{i,t} - \frac{1}{|G_{g(i),t}|} \sum_{j \in G_{g(i),t}} \hat{y}_{j,t}, \quad (3.4)$$

$$\tilde{y}_{i,t} = y_{i,t} - \frac{1}{|G_{g(i),t}|} \sum_{j \in G_{g(i),t}} y_{j,t}, \quad (3.5)$$

其中 $G_{g(i),t}$ 為日 t 時產業 $g(i)$ 的有效股票集合（排除標籤為 NaN 者）。產業中性 IC 定義為

$$IC_t^{\text{ind-neutral}} = \text{Corr}(\tilde{\hat{y}}_{\cdot,t}, \tilde{y}_{\cdot,t}). \quad (3.6)$$

誤差指標 (Error metrics): 以 MSE、RMSE 與 MAE 衡量預測誤差。

3.5.2 投資組合績效指標

年化報酬率 (Annualized return): 以每次再平衡選取預測分數最高之 top_pct 股票等權做多，計算截面平均報酬作為策略報酬，並依再平衡頻率年化。

Sharpe Ratio: 以平均報酬除以波動後再乘年化因子，衡量風險調整後報酬。

勝率 (Hit ratio): 正報酬期間比例，用以衡量策略穩定性。

3.5.3 基準比較與流程摘要

天真基準 (Naive baseline): 令所有股票預測值為 0，計算 MSE/RMSE/MAE 作為誤差下界參考。

台灣 50 ETF 基準: 以 0050 之 forward- k 日報酬作為市場基準，與策略累積報酬比較。

評估流程為：載入 artifacts 與模型權重後，分別於訓練期與測試期計算預測與投資組合指標；並輸出 Daily IC、IC 分佈與累積報酬等視覺化結果，以利比較不同模型之表現。對於 DMFM，另輸出注意力權重分佈作為解釋性分析依據。

第四章、實驗結果

4.1 模型對比

(此處撰寫模型對比內容)

4.2 投組績效驗證

(此處撰寫投組績效驗證內容)

4.3 特徵重要性排名

(此處撰寫特徵重要性排名內容)

第五章、結論

5.1 主要發現

(此處撰寫主要發現內容)

5.2 與既有研究的對比

(此處撰寫與既有研究的對比內容)

5.3 模型限制

(此處撰寫模型限制內容)

5.4 實踐應用與未來方向

(此處撰寫實踐應用與未來方向內容)

參考文獻

- [1] M. Agrawal, P. K. Shukla, R. Nair, A. Nayyar, and M. Masud, “Stock prediction based on technical indicators using deep learning model,” *Computers, Materials & Continua*, vol. 70, no. 1, pp. 287–304, 2022.
- [2] M. Agrawal, A. U. Khan, and P. K. Shukla, “Stock price prediction using technical indicators: A predictive model using optimal deep learning,” *International Journal of Recent Technology and Engineering (IJRTE)*, vol. 8, no. 2, pp. 2297–2305, 2019.
- [3] A. W. Li and G. S. Bastos, “Stock market forecasting using deep learning and technical analysis: A systematic review,” *IEEE Access*, vol. 8, pp. 185 197–185 215, 2020.
- [4] S. Zaheer, N. Anjum, S. Hussain, A. D. Algarni, J. Iqbal, S. Bourouis, and S. S. Ullah, “A multi parameter forecasting for stock time series data using lstm and deep learning model,” *Mathematics*, vol. 11, no. 3, p. 590, 2023.
- [5] J. Qiu, B. Wang, and C. Zhou, “Forecasting stock prices with long-short term memory neural network based on attention mechanism,” *PLoS ONE*, vol. 15, no. 1, p. e0227222, 2020.
- [6] 鄭邦廷, “基於深度學習與技術分析指標預測股市買賣點,” Master’s thesis, 國立臺灣師範大學機電工程學系, 台北, 台灣, 2023, stock buy and sell points prediction based on deep learning and technical analysis indicators.

- [7] 廖俊翔, “應用深度學習結合自相關分析優化股票預測模型,” Master’s thesis, 國立中興大學資訊管理學系, 台中, 台灣, 2022, applying deep learning combined with autocorrelation analysis to optimize stock forecasting model.
- [8] Z. Wu, S. Pan, F. Chen, G. Long, C. Zhang, and P. S. Yu, “A comprehensive survey on graph neural networks,” *IEEE Transactions on Neural Networks and Learning Systems*, vol. 31, no. 1, pp. 4–24, 2020.
- [9] K. Huang, X. Li, F. Liu, X. Yang, and W. Yu, “Ml-gat: A multilevel graph attention model for stock prediction,” *IEEE Access*, vol. 10, pp. 85 472–85 483, 2022.
- [10] G. Song, T. Zhao, S. Wang, H. Wang, and X. Li, “Stock ranking prediction using a graph aggregation network based on stock price and stock relationship information,” *Information Sciences*, vol. 643, p. 119236, 2023.
- [11] L. Cheng, J. Wen, and Y. Wang, “Stock prediction model based on multi-feature graph attention network,” *Nanjing University of Aeronautics and Astronautics (Manuscript)*, 2024, manuscript.
- [12] Z. Wei, B. Dai, and D. Lin, “Factor investing with a deep multi-factor model,” *arXiv preprint*, 2022, under Review.

Appendix 附錄

本附錄彙整本研究模型輸入特徵之完整清單。

Table 3: 本研究實際使用之模型輸入特徵 (feature_cols, 共 54 項；全部由 TEJ 原始欄位衍生)

特徵名稱	類別	定義（由 TEJ 欄位計算）	視窗
ret_1	報酬	$P_t/P_{t-1} - 1$, 其中 P_t 為 TEJ 收盤價(元)	1
ret_3	報酬	$P_t/P_{t-3} - 1$ (TEJ 收盤價)	3
ret_5	報酬	$P_t/P_{t-5} - 1$ (TEJ 收盤價)	5
ret_10	報酬	$P_t/P_{t-10} - 1$ (TEJ 收盤價)	10
ret_20	報酬	$P_t/P_{t-20} - 1$ (TEJ 收盤價)	20
rev_1	反轉	$-ret_1$ (由 TEJ 收盤價衍生)	1
rev_5	反轉	$-ret_5$ (由 TEJ 收盤價衍生)	5
rev_10	反轉	$-ret_10$ (由 TEJ 收盤價衍生)	10
px_over_sma_5	趨勢	$P_t/SMA_5(P)_t - 1$ (TEJ 收盤價)	5
px_over_sma_10	趨勢	$P_t/SMA_{10}(P)_t - 1$ (TEJ 收盤價)	10
px_over_sma_20	趨勢	$P_t/SMA_{20}(P)_t - 1$ (TEJ 收盤價)	20
px_over_sma_60	趨勢	$P_t/SMA_{60}(P)_t - 1$ (TEJ 收盤價)	60
mom_diff_10	動能差	$ret_10 - ret_1$ (由 TEJ 收盤價衍生)	-

(續下頁)

特徵名稱	類別	定義（由 TEJ 欄位計算）	視窗
mom_diff_20	動能差	$\text{ret}_{-20} - \text{ret}_{-1}$ (由 TEJ 收盤價衍生)	-
std_ret_5	波動	$\text{Std}(r_{t-4}, \dots, r_t)$, r_t 由 TEJ 收盤價計算	5
std_ret_10	波動	$\text{Std}(r_{t-9}, \dots, r_t)$ (TEJ 收盤價)	10
std_ret_20	波動	$\text{Std}(r_{t-19}, \dots, r_t)$ (TEJ 收盤價)	20
std_ret_60	波動	$\text{Std}(r_{t-59}, \dots, r_t)$ (TEJ 收盤價)	60
skew_20	分配	20 日報酬偏度 (由 TEJ 收盤價計算之 r_t)	20
kurt_20	分配	20 日報酬峰度 (由 TEJ 收盤價計算之 r_t)	20
vol_over_ma_5	量能	$V_t / \text{SMA}_5(V)_t - 1$, V_t 為 TEJ 成交量 (千股)	5
vol_over_ma_10	量能	$V_t / \text{SMA}_{10}(V)_t - 1$ (TEJ 成交量)	10
vol_over_ma_20	量能	$V_t / \text{SMA}_{20}(V)_t - 1$ (TEJ 成交量)	20
vol_over_ma_60	量能	$V_t / \text{SMA}_{60}(V)_t - 1$ (TEJ 成交量)	60
up_vol_20	量能	20 日內上漲日平均量: $\frac{\sum V \mathbb{I}(r>0)}{\sum \mathbb{I}(r>0)+\varepsilon}$ (TEJ 成交量、TEJ 收盤價計算 r)	20
down_vol_20	量能	20 日內下跌日平均量: $\frac{\sum V \mathbb{I}(r<0)}{\sum \mathbb{I}(r<0)+\varepsilon}$ (TEJ 成交量、TEJ 收盤價計算 r)	20
amihud_5	流動性	Amihud illiquidity: $\frac{1}{5} \sum \frac{ r }{\text{Amt}+\varepsilon}$, Amt 為 TEJ 成交值 (千元)	5
amihud_20	流動性	同上 (TEJ 成交值、TEJ 收盤價計算 r)	20

(續下頁)

特徵名稱	類別	定義（由 TEJ 欄位計算）	視窗
rsi_14	技術指標	RSI (TA-Lib 預設參數；由 TEJ 收盤價計算)	14
stoch_k_14	技術指標	Stochastic %K (TA-Lib 預設；由 TEJ 最高/最低/收盤價計算)	14
stoch_d_3	技術指標	Stochastic %D (%K 的平滑；TA-Lib 預設)	3
macd	技術指標	MACD 主線 (TA-Lib 預設；由 TEJ 收盤價計算)	-
macd_signal	技術指標	MACD 訊號線 (TA-Lib 預設)	-
macd_hist	技術指標	MACD 柱狀差 (macd-macd_signal)	-
atr_14	風險	ATR (Average True Range；由 TEJ 最高/最低/收盤價計算)	14
mdd_20	風險	20 日最大回撤： $\max_{\tau \leq t} \left(\frac{\max P - P}{\max P} \right)$ (TEJ 收盤價)	20
beta_60	風險	60 日 rolling beta (由 TEJ 資料計算)：以個股日報酬 $r_{i,t}$ 對市場日報酬 $r_{m,t}$ 做 OLS， $\beta = \frac{\text{Cov}(r_i, r_m)}{\text{Var}(r_m)}$ 。本研究之市場報酬 $r_{m,t}$ 使用 TEJ 匯出之 0050 (證券代碼 0050) 收盤價計算日報酬。	60

(續下頁)

特徵名稱	類別	定義 (由 TEJ 欄位計算)	視窗
idio_vol_60	風險	60 日特質波動：上述 rolling 回歸殘差 $\varepsilon_{i,t}$ 的標準差 $\text{Std}(\varepsilon)$ 。回歸所用市場報酬同 beta_60，亦由 TEJ 匯出之 0050 收盤價計算。	60
roll_max_5	區間極值	5 日 rolling maximum: $\max(P_{t-4}, \dots, P_t)$ (TEJ 收盤價)	5
roll_min_5	區間極值	5 日 rolling minimum: $\min(P_{t-4}, \dots, P_t)$ (TEJ 收盤價)	5
pct_pos_5	區間位置	$\frac{P_t - \text{roll_min_5}}{\text{roll_max_5} - \text{roll_min_5} + \varepsilon}$ (TEJ 收盤價)	5
roll_max_10	區間極值	10 日 rolling maximum (TEJ 收盤價)	10
roll_min_10	區間極值	10 日 rolling minimum (TEJ 收盤價)	10
pct_pos_10	區間位置	$\frac{P_t - \text{roll_min_10}}{\text{roll_max_10} - \text{roll_min_10} + \varepsilon}$ (TEJ 收盤價)	10
roll_max_20	區間極值	20 日 rolling maximum (TEJ 收盤價)	20
roll_min_20	區間極值	20 日 rolling minimum (TEJ 收盤價)	20
pct_pos_20	區間位置	$\frac{P_t - \text{roll_min_20}}{\text{roll_max_20} - \text{roll_min_20} + \varepsilon}$ (TEJ 收盤價)	20
roll_max_60	區間極值	60 日 rolling maximum (TEJ 收盤價)	60
roll_min_60	區間極值	60 日 rolling minimum (TEJ 收盤價)	60
pct_pos_60	區間位置	$\frac{P_t - \text{roll_min_60}}{\text{roll_max_60} - \text{roll_min_60} + \varepsilon}$ (TEJ 收盤價)	60

(續下頁)

特徵名稱	類別	定義（由 TEJ 欄位計算）	視窗
pct_to_high_20	位置差距	距 20 日高點比例: $\frac{P_t}{\text{roll_max_20}} - 1$ (TEJ 20 收盤價)	
pct_to_low_20	位置差距	距 20 日低點比例: $\frac{P_t}{\text{roll_min_20}} - 1$ (TEJ 20 收盤價)	
zscore_close_20	標準化	20 日 z-score: $\frac{P_t - \mu_{20}}{\sigma_{20} + \varepsilon}$ (TEJ 收盤價; μ, σ 20 為 rolling mean/std)	
zscore_close_60	標準化	60 日 z-score: $\frac{P_t - \mu_{60}}{\sigma_{60} + \varepsilon}$ (TEJ 收盤價)	60

