AI CUP 2024 秋季賽
根據區域微氣候資料預測發電量競賽報告

隊伍：TEAM_6668
隊員：王禮芳（隊長）、莊秉宸、李旻昊、余振揚
Private leaderboard：386003.4 / Rank 4

壹、環境

- **作業系統**：Ubuntu 18.04.6LTS
- **程式語言**：Python 3.9
- **CPU:** Intel Xeon Gold 6138 20C 2.0GHz * 2（未使用 GPU）
- **Packages（如 Github Requirements.txt）：**
    - sklearn
    - optuna
    - numpy
    - pandas

- **預訓練模型**：
  本研究使用機器學習模型，未使用額外預訓練之深度學習模型。

- **額外資料集：**
  在整理資料集時，為獲得更多額外的氣象資訊，我們參考了中央氣象局之氣候觀測資料查詢服務[1]，取得東華站站點之全天空日射量與每日降水量，以及氣象局所公布之臺灣各地四季太陽仰角與方位角文件[2]，以緯度與花蓮相近的台中作為太陽仰角與方位角的特徵。

  以下是本研究所蒐集之額外資料集：
    - 東華站 (C0Z100) 的全天空日射量
    - 東華站 (C0Z100) 的每日降水量
    - 臺灣各地四季太陽仰角與方位角

貳、演算方法與模型架構

一、演算方法

在這次的預測發電量任務中，我們將目標測試集拆分成兩百個子問題，建立各自的子迴歸模型。我們使用欲預測時段正在運作的太陽能發電站資料，與目標發電站的歷史資料取交集作為訓練資料，並建立聚合型的的隨機森林模型進行訓練，藉此預測出指定發電站在當日特定時段的發電量數據。

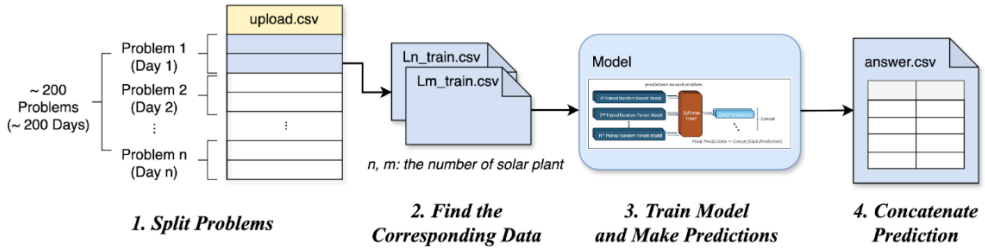
本研究的演算方法能以一四階段的 pipeline 表示（如圖一）：

1. **切分子問題**
   將目標預測任務以日為單位，切分成兩百個子問題（一天視為一個問題），並建立「目標時間和站點-運轉中發電站」字典。

2. **尋找對應的訓練資料**
   我們透過篩選目標發電站與預測日期同時運轉的其他發電站，取得其歷史資料的交集。將目標發電站的發電量視為標籤值，並將預測日期同時運轉的其他發電站的相關資料作為特徵，構建模型訓練資料集。

3. **模型訓練及預測**
   我們為每個目標發電站與目標日期中運轉的發電站建立隨機森林模型，並將預測輸出通過 Softmax 層，根據模型 MAE 分配權重，將預測發電量進行加權聚合。

4. **對所有子題進行迭代連接**
   以上流程將分別對兩百個子問題逐一進行，並將每個迭代出的預測值進行串接，生成最終的輸出檔案。



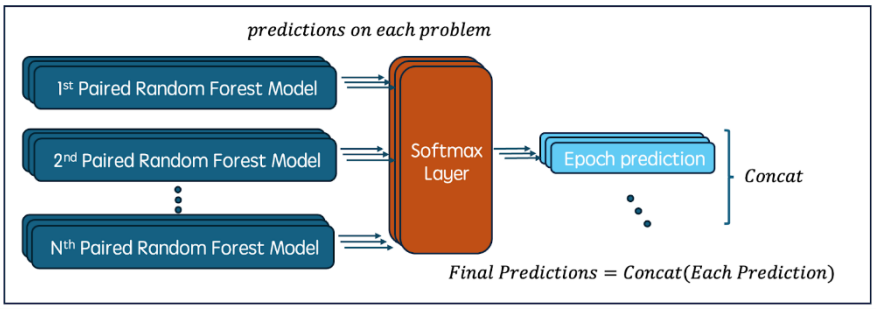
圖一、演算流程

二、模型架構與參數設定

本研究以多個隨機森林模型聚合發電量預測值（如圖二），主要包含兩部分：

1. **成對的隨機森林模型**
   我們以目標發電站的發電量為標籤，利用同時也在運轉的另一發電站其他特徵進行訓練，建置出一由兩成對的發電站資料訓練而成的隨機森林模型。

2. **Softmax 層**
   在合併各隨機森林模型時，我們引入 softmax 層為各模型預測值加上權重，一模型的 MAE 愈小，將給其愈大的權重值。換句話說，我們期待能由此提升誤差較小的模型之影響力，同時亦想保留由其他發電站特徵訓練出的模型所挾帶的資訊。



圖二、模型架構示意圖

在最終的上傳階段，本研究使用之各項參數設定如下：

訓練參數	
n_estimator	100
Softmax Beta	0.1
criterion	gini
min_samples_split	2
min_samples_leaf	1

表一、各項訓練參數

參、創新性

一、將全局問題拆分成 200 個子問題

考量到每個發電站具有其獨特性（如高度、面向角度），這些因素會對日照量與發電量產生顯著影響，若採用全局模型，難以達到個別化調整。因此，我們將問題劃分為 200 個子問題，將目標站點的發電量作為標籤值，並將目標日運作中的其他站點資訊作為訓練特徵。透過篩選歷史資料的交集，訓練模型以捕捉站點間的線性與非線性關係，**將複雜的全局預測問題分解為兩個站點間的迴歸任務。**

二、利用當日運轉發電站資料，預測當日目標發電站發電量

根據競賽官方文件，由於光照度感測器有感測的最大數值限制，因此當光照度到達該上限值時，或許將不再是預測發電量的可靠特徵。為了處理感測器缺陷補償，本研究利用 XGBoost 模型，以未到達上限的光照度為標籤訓練模型，對到達上限的光照度進行迴歸，填補當時可能的真實光照度。

三、聚合針對每個目標天輸出的 Softmax 層

在特定目標日中，可能會有多個正在運行的感測器。在我們的訓練流程中，會針對目標站和每個目標天運行的感測器模型，分別建立對應的模型。考慮到不同站點可能含有其隱含的特性，**我們引入了基於模型 MAE 進行加權的 Softmax 層**，對各個模型於該日的預測光照亮進行加權聚合，生成最終的輸出值。

肆、資料處理

一、資料集切分

由於每日的上傳次數限制，我們按照 upload.csv 格式，對提供的訓練資料切分出訓練集和測試集，方便進行不同方法和超參數遍歷的效能評估。

二、時序轉換和資料增強

1. 考量到分鐘為單位的資料集數量巨大，我們將資料以十分鐘為區段進行平均，以減少資料量。

2. 由於時間的 Datetime 格式無法作為訓練特徵，我們提取時間資訊並建立了以下四項時間相關資訊：
    - [Day]：表示目標日與當年 1 月 1 日的相差天數；
    - [Month]：目標日期的月份；
    - [Hour]：目標時間的小時數；
    - [Minute]：目標時間的分鐘數。

3. 考量到所使用的 tree-based 模型將每一列資料獨立處理，會使模型無法捕捉列間的時序關係，因此我們增加了滯後特徵，包括前十分鐘的發電量 [lag_power] 和日照量 [lag_sunlight]，使模型學習到歷史資訊的時序性。

4. 由於日照量和發電量間存在超過 70% 的相關性，我們加入了一項交互作用特徵 [sun_power_inter] = sunlight * power，捕捉兩特徵間的非線性關係。

三、目標測試集的填補缺失值

在我們的方法中，我們針對目標天找到該天正在運作的裝置，並將該站所紀錄的資訊用於輸出的特徵值。然而並不是每個站點都含有完整 9:00 ~ 16:50 的特徵紀錄，因此我們使用向前填補和向後填補特徵的方式，建立目標天中完整時序的特徵矩陣。

而針對目標天無其他站點運作中的情形，我們使用目標裝置在該時間 (e.g. 0900) 裡最相近的四天發電量平均，作為目標天的發電量輸出，詳細方法將於訓練方式段落中提出。

四、額外資料集

考量額外的氣候特徵可能對於太陽能發電有一定程度的影響，本研究在資料蒐集的初期，納入了額外的氣候相關特徵。

我們蒐集了更多全局性的資料特徵，例如來自東華氣象站的全天空日射量與降水量[1]，以及與花蓮緯度相近的台中之太陽仰角與方位角[2]。同時，我們也計算了主辦單位所提供之各發電站點的太陽能版面朝方向與太陽方位角的餘弦相似度，作為補充的特徵。

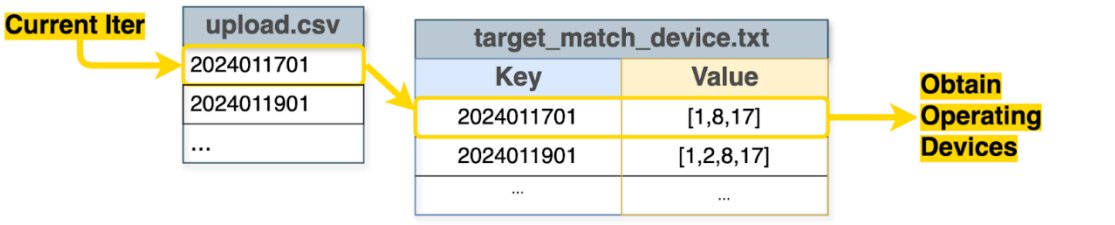
伍、訓練方式

在這次的任務中，我們分成四個階段進行：第一階段中，我們首先建立了「時間-運轉中發電站」的 mapping table 。第二階段我們針對目標發電站和目標日運作中發電站，提取其歷史時間交集的資料，作為模型訓練的依據。第三階段中，我們在目標日的每個站點上分別訓練隨機森林模型，並通過以 MAE 為權重依據的 Softmax 層，進行發電量預測值的加權聚合，第四階段則對拆分的 200 個問題逐一進行預測，最終將所有預測的發電量連接成完整的輸出檔案。

一、建立一日期-運轉中發電站的 mapping table

透過觀察訓練資料集，我們發現大部分的目標時間在其他站點均有特徵和發電紀錄，因此我們將資料集切分成 200 天，並建立一根據題目的「時間-運轉中發電站」mapping table 字典。其中 key 表示為目標日期和裝置，value 為目標裝置和當日有運轉紀錄的發電站列表。

以 2024/1/17 為例，由於裝置 8 和裝置 17 在 1/17 皆有發電紀錄和當日天氣特徵，因此我們建立的 mapping table 為 { "2024011701": [1, 8, 17] }。在未來使用時，便能查表找出特定日期運轉中的發電站點（如圖三）。



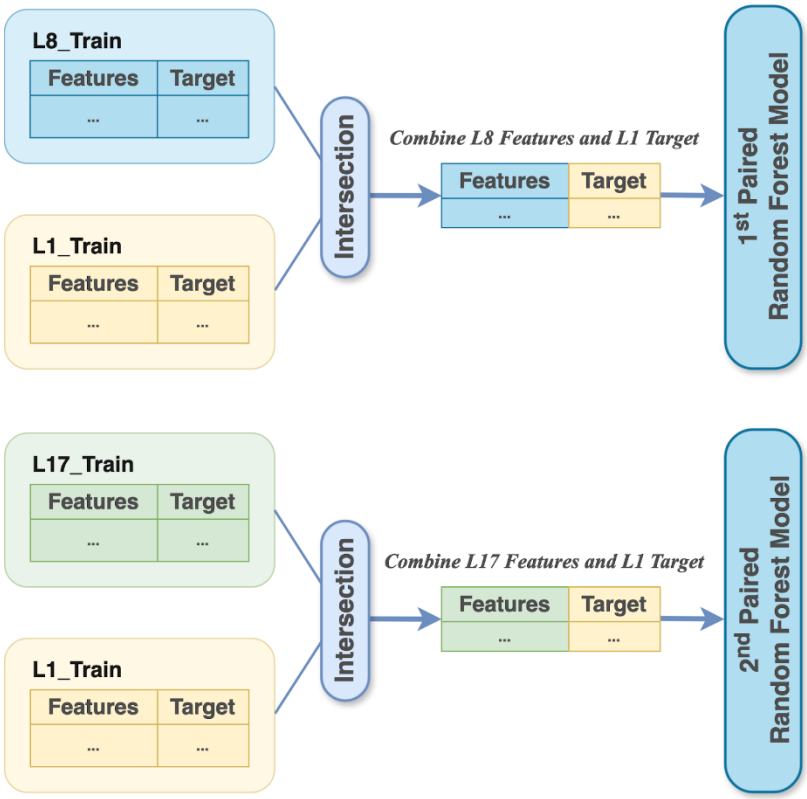
圖三、時間-運轉中發電站 mapping table 之圖例

二、準備訓練資料

在取得 mapping table 後，我們將針對目標發電站，以及目標日運作發電站，取得兩者在時間上重疊的歷史資料。而後將目標發電站的發電量作為標籤值，而目標日運作發電站的天氣資訊和發電量將做為訓練使用的特徵。

例如，在預測 2024/1/17 之一號發電站於 9:00~16:50 期間每十分鐘的平均發電量時，我們將分別利用 8 號和 17 號發電站的資料，與目標 1 號發電機的資料個別取交集。而後以 8 號和 17 號發電機的各項特徵作為模型輸入，目標一號發電機的發電量作為標記，整理出「8 號對 1 號」、「17 號對 1 號」兩個資料集，並分別用於訓練兩個模型（如圖四）。

需要注意的是，由於在題目指定的時段中一號發電機沒有發電資料，因此該時段將不被納入訓練過程。



圖四、以與目標發電機的配對資料送入模型示意圖

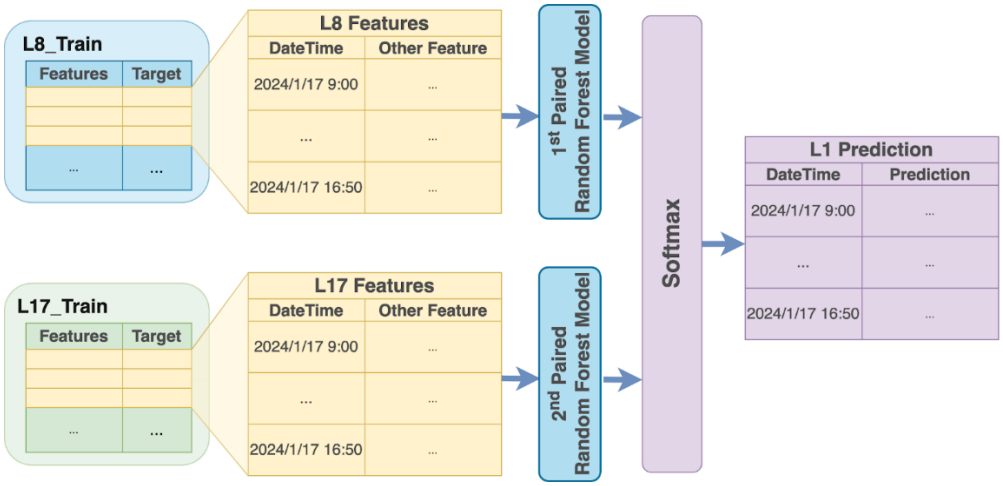
三、訓練 Softmax 聚合的隨機森林模型

此時，我們將利用上一步驟所得的訓練資料，針對每個目標日運轉中的發電站建構隨機森林模型。具體來說，若多個發電站在目標預測日裡都有運轉記錄，則當日將會依據這些發電站的數量，分別建立對應數量的隨機森林模型進行訓練。

由此，我們會得到 N 個訓練好的隨機森林模型，並且在輸入該站點於目標日的測試集特徵後（包含從 9:00 至 16:50，共 48 筆資料），獲得長度為 N*48 的預測發電量矩陣。而後，我們將此發電量矩陣送入基於 MAE 調整權重的 Softmax 層（公式如下），加權聚合成長度為 1*48 的預測向量。

$$Weight_i = \frac{\exp(-\beta * MAE_i)}{\sum_j \exp(-\beta * MAE_j)}$$

例如，針對 2024/1/17 時的 1 號發電機發電量預測任務，我們已分別針對 8 號發電機與 17 號訓練出了預測 1 號發電機的模型。預測過程中，我們以兩發電站 2024/1/17 9:00 至 16:50 間每十分鐘特徵為輸入，預測一號發電機於各時段的發電量，最終經由一 Softmax 層進行加權計算，得到最終的預測值（如圖五）。



圖五、模型預測示意圖

四、對 200 個問題進行預測連接

在模型預測過程中，將反覆迭代上述步驟，直到 200 個子問題皆計算完畢。
最後我們將所有子問題的預測結果串接，合併為最終上傳系統的預測發電量檔案。

$$Final\ Prediction\ =\ Concat(Prediction\ of\ each\ problem)$$

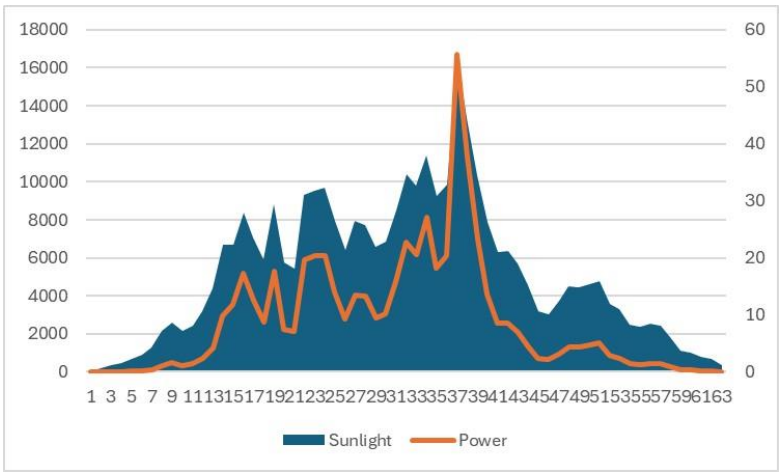
然而，有時會發生目標日裡沒有其他發電站運轉的情形。此時我們將找出目標發電站在目標時段（如 9:00）裡，最接近目標日的任意四天資料，計算出發電量的平均值。而最接近目標日的四天，指計算所有資料日期與目標日相隔的天數，取前四筆最接近者，不限於往前或往後取值。

陸、分析與結論

一、各面向分析

1. 光照、發電量分析

本研究先分析各類特徵對發電量的趨勢，發現光照值對發電量有顯著的相關性。將兩者趨勢疊圖，以每日時間點 ID 1 ~ 63 為 x 軸，我們能觀察到兩者的緊密關聯（如圖六）：



圖六、日照與發電量趨勢圖

2. 外部資料分析

在資料增強部分，我們認為加入額外的氣候資訊可能有助於訓練，因此蒐集了具有全局性的資料特徵（如東華氣象站的全天空日照量、降雨量、台中的太陽仰角等等資訊）。

然而在實驗中我們看到了這些額外尋找的全局性特徵並無助於模型的學習，甚至不考慮這些額外特徵，能使模型表現得更好。因此，在最後的訓練中，我們最終移除了這些額外的氣候資訊。

在我們最終採用的隨機森林演算法中，每棵決策樹的訓練僅考慮各列資料本身的特徵，而不考慮列資料間的相互和時序依賴關係，即資料間是彼此獨立的。這會使得我們加入的額外資料成為靜態特徵，具有極低的特徵重要度。因此，我們認為雖然這些資料帶來了額外的物理與天氣資訊，但在我們的模型中難以發揮作用。

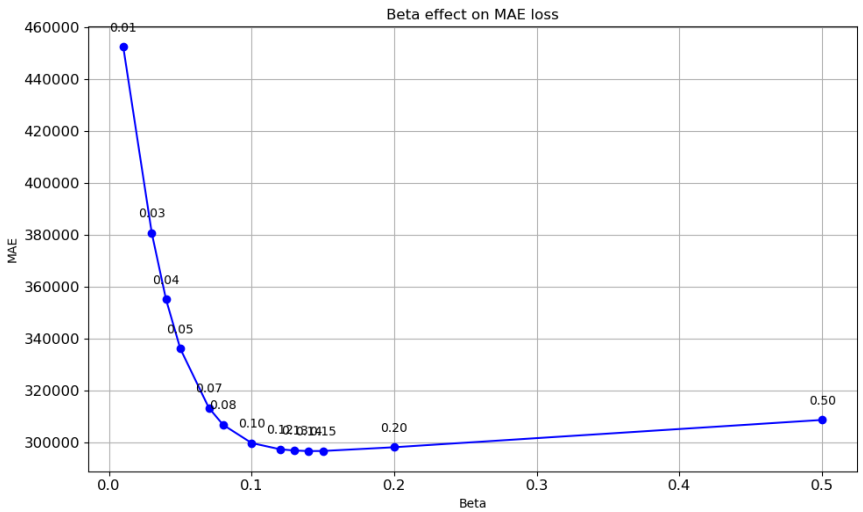
3. Softmax beta 值分析

在聚合預測值的 Softmax 層中（公式如下），我們使用 Beta 值來調整模型間的權重影響力，即較大的 Beta 值會使 MAE 較小的模型有更大的權重，而 MAE 較大的模型效果會被稀釋。

$$Weight_i = \frac{\exp(-\beta * MAE_i)}{\sum_j \exp(-\beta * MAE_j)}$$

我們在自行切割的資料集上進行不同 Beta 值對模型表現的影響分析（如圖七），能發現 beta 約在 0.12~0.13 時能使模型有更優秀的表現，這是因為此時的權重分配適度強化了 MAE 較低的模型，而未過度削弱 MAE 較高的模型。

MAE 較低的模型通常代表目標裝置與該裝置在靜態特徵（如高度、面向角度等）上具有更高的相似性。因此在較大的聚合權重下，能更貼近的反映出目標站點的資料特性。然而，若 Beta 值過大，過度強調了 MAE 低的模型，忽略了其他站點的資訊，會降低模型整體的表現。



圖七、SoftMax beta 值對模型 Loss 折線圖

二、改進方向與未來展望

1. 加入交叉驗證

在本次競賽過程中，我們並未引入交叉驗證。鑑於交叉驗證是一種選擇最佳參數與評估模型泛化能力的重要技術，我們認為未來若能在訓練過程中加入交叉驗證，或能有助於更精確地評估模型的穩定性與性能表現，並減少模型過擬合的風險。

2. 未遍歷所有超參數組合

本研究在選定超參數時，並未遍歷所有可能組合。若在未來要進一步研究相關主題，應更加嚴謹地選定最適當的一組超參數，期待能由此提升模型預測能力。若在未來的研究中能借助更高效的參數搜尋方法（如貝葉斯優化或遺傳算法），將能更系統地探索超參數空間，並嚴謹選擇最適當的超參數組合，期待藉此進一步提升模型的預測準確性。

3. 使用 stacking 方法

雖在研究初期，本隊伍曾嘗試使用 stacking 技巧，結合多種機器學習模型進行集成訓練，但因模型間的協同效果在短期內未能達到預期而暫時放棄。然而，我們依然認為 stacking 是一有潛力的策略，期望能在未來進行相關研究時，進行更進一步的 stacking 技術實作，如調整模型權重、引入更多元的模型等等，以此發揮集成學習的優勢，進一步提升模型的表現。

4. 加入動態特徵

在本次競賽中，我們所蒐集的額外特徵資料皆是靜態的、全局性的，例如如東華氣象站的全天空日照量。但是這樣的資訊在本次競賽裡，並無法有效地幫助模型提升預測能力，因此我們認為，尋找更多動態的局部特徵，或許也能幫助模型提升表現。

柒、程式碼

Github 連結：https://github.com/Mere-cat/AI-CUP-2024/tree/main

環境設置：

pip install -r requirements.txt

程式執行：

執行 BlaBlaBlazzz 目錄裡的 run.py

python3 run.py

捌、使用的外部資源與參考文獻

本隊伍於此次比賽期間，撰寫程式時，部分經由 ChatGPT 協助除錯。

[1] 交通部中央氣象署。 (n.d.) 。資料瀏覽。 CODIS 氣候觀測資料查詢服務。 https://codis.cwa.gov.tw/StationData

[2] 交通部中央氣象署。 (n.d.) 。臺灣四季太陽仰角與方位角。 https://www.cwa.gov.tw/Data/astronomy/season.pdf

作者聯絡資料表

隊伍名稱	TEAM6668	Private Leaderboard 成績	386003.4	Private Leaderboard 名次	4
身分 （隊長/隊員）	姓名 (中英皆需填寫) (英文寫法為名, 姓，例：Xiao－Ming, Wu，名須加連字號，姓前須加逗號)	學校＋系所中文全稱 (請填寫完整全名，勿縮寫)	學校＋系所英文中文全稱 (請填寫完整全名，勿縮寫)	電話	E-mail
隊長	王禮芳 Li-Fang, Wang	國立成功大學敏求智慧運算學院	National Cheng Kung University, Miin Wu School of Computing	0963-493-999	nm6134059@gs.ncku.edu.tw
隊員 1	莊秉宸 Ping-Cheng, Chuang	國立成功大學統計學系	National Cheng Kung University, Department of Statistics	0979-843-319	jason20031215@gmail.com
隊員 2	余振揚 Chen-Yang Yu	國立成功大學敏求智慧運算學院	National Cheng Kung University, Miin Wu School of Computing	0902-231-856	nm6121030@gs.ncku.edu.tw
隊員 3	李旻昊 Man-Ho, Li	國立成功大學敏求智慧運算學院	National Cheng Kung University, Miin Wu School of Computing	0989-852-027	limarco310@gmail.com
隊員 4					

★註 1：請確認上述資料與 AI CUP 報名系統中填寫之內容相同。自 2023 年起，獎狀製作將依據報名系統中填寫內容為準，有特殊狀況需修正者，請主動於報告繳交期限內來信 moe.ai.ncu@gmail.com。報告繳交截止時間後將不予修改。

★註 2：繳交程式碼檔案與報告，請 Email 至：ailabailab5051@gmail.com，並同時副本至：t_brain@trendmicro.com 與 moe.ai.ncu@gmail.com。缺一不可。