國 立 中 央 大 學

資訊工程學系資訊工程碩士班

碩士論文

結合時空資料的半監督模型並應用於 PM2.5 空污感測器的異常偵測

Semi-Supervised Model with Spatio-Temporal Data and Applied in PM2.5 sensor anomaly detection

研 究 生：張欣茹

指導教授：陳弘軒　博士

中 華 民 國 一百一十 年 六 月

國立中央大學圖書館學位論文授權書

填單日期：110 / 8 / 5　　2019.9 版

授權人姓名	張玟茹.	學 號	108522084.
系所名稱	資訊工程學系.	學位類別	☑碩士　□博士
論文名稱	結合時空資料的半監督模型並應用於PM2.5室內感測器的異常偵測.	指導教授	陳弘軒　教授.

學位論文網路公開授權

授權本人撰寫之學位論文全文電子檔：

・在「**國立中央大學圖書館博碩士論文系統**」.

(✓)同意立即網路公開

(　)同意 於西元＿＿＿年＿＿＿月＿＿＿日網路公開

(　)不同意網路公開，原因是：＿＿＿＿＿＿＿＿＿＿

・在**國家圖書館「臺灣博碩士論文知識加值系統」**

(✓)同意立即網路公開

(　)同意 於西元＿＿＿年＿＿＿月＿＿＿日網路公開

(　)不同意網路公開，原因是：＿＿＿＿＿＿＿＿＿＿

依著作權法規定，非專屬、無償授權國立中央大學、台灣聯合大學系統與國家圖書館，不限地域、時間與次數，以文件、錄影帶、錄音帶、光碟、微縮、數位化或其他方式將上列授權標的基於非營利目的進行重製。

學位論文紙本延後公開申請 (紙本學位論文**立即公開者此欄免填**)

本人撰寫之學位論文紙本因以下原因將延後公開

・**延後原因**

(　)已申請專利並檢附證明，專利申請案號：

(　)準備以上列論文投稿期刊

(　)涉國家機密

(　)依法不得提供，請說明：＿＿＿＿＿＿＿＿

・**公開日期**：西元＿＿＿＿年＿＿＿＿月＿＿＿＿日

※繳交教務處註冊組之紙本論文(送繳國家圖書館)若不立即公開，請加填「國家圖書館學位論文延後公開申請書」

研究生簽名：張玟茹.　　指導教授簽名：陳弘軒

*本授權書請完整填寫並親筆簽名後，裝訂於論文**封面之次頁**。

國立中央大學碩士班研究生
論 文 指 導 教 授 推 薦 書

資訊工程學系碩士班 學系/研究所 張欣茹 研究生

所提之論文 結合時空資料的半監督模型並應用於PM2.5空污感測器的異常偵測

係由本人指導撰述，同意提付審查。

指導教授 _____ (簽章)

110 年 7 月 12 日

1100711

國立中央大學碩士班研究生
論 文 口 試 委 員 審 定 書

資訊工程學系碩士班 學系/研究所 張欣茹 研究生

所提之論文 結合時空資料的半監督模型並應用於PM2.5空污感測器的異常偵測

經由委員會審議，認定符合碩士資格標準。

學位考試委員會召集人 _____

委　　員 _____

_____

_____

_____

中　華　民　國　110　年　7　月　26　日

1100720

國立中央大學碩士班研究生
論 文 口 試 委 員 審 定 書

資訊工程學系碩士班 學系/研究所 張欣茹 研究生

所提之論文 結合時空資料的半監督模型並應用於PM2.5空污感測器的異常偵測

經由委員會審議，認定符合碩士資格標準。

學位考試委員會召集人 _____

委　　員 _____

_____

_____

_____

中　華　民　國　　年　　月　　日

1100720

NameService

結合時空資料的半監督模型並應用於 PM2.5 空污感測器的異常偵測

摘要

台灣近年來 PM2.5 空氣汙染的議題逐漸受到重視，增設了許多價格較為低廉的感測器，但是這些感測器容易受到環境因素影響造成較大的誤差，加上數量龐大造成每台感測器的維護頻率低，單一區域感測器回傳的數值不如國家級測站來得可靠，

本論文比較了監督式、無監督式、及半監督式的演算法在偵測異常傳感器的效果。為了結合感測器的時空資訊，我們將監測值轉成圖片資料、整合性資料、以及整合資料結合時序資料來準備訓練數據。我們根據工業技術研究所提供的檢測記錄得到感器測的狀態值（正常或異常），探討了標記資料的比例對半監督模型預測效能的影響。實驗結果顯示：我們研究的方法優於目前的隨機巡檢機制。

關鍵字： PM2.5, 異常偵測, 半監督模型, 時空資料結合

Semi-Supervised Model with Spatio-Temporal Data and Applied in PM2.5 sensor anomaly detection

Abstract

The PM2.5 issue has drawn much attention in Taiwan, and many inexpensive sensors have been deployed in recent years. However, these sensors are fragile and susceptible to environmental factors. In addition, the large number of sensors results in low maintenance frequency, so the monitored values returned by a single sensor are unreliable.

This thesis compares supervised, unsupervised, and semi-supervised methods to identify the problematic sensors. We prepared the training data by converting monitored values into images, integrated data, and sequential data to incorporate the spatio-temporal information of the sensors. We obtained sensors' status (normal or abnormal) based on the inspection records provided by the Industrial Technology Research Institute. We explored how the ratio of labeled data to unlabeled data influences the performance of the semi-supervised models. Experimental results show that our studied methods outperform the current inspection strategy (random inspection).

Keywords: PM2.5, anomaly detection, semi-supervised model, spatio-temporal data integration

目錄

圖目錄

頁次

表目錄

頁次

一、 緒論

1.1 研究動機

台灣近年的環保意識抬頭，空氣汙染及 PM2.5 的問題逐漸為國人重視。PM2.5 指的是粒徑小於 2.5 微米 (μm) 的懸浮微粒，因 PM2.5 的體積極小，人體的黏膜或纖毛無法阻隔其進入人體，且 PM2.5 的成份複雜，可能附著著對人體有害的物質，故長期接觸 PM2.5 可能致癌，世界衛生組織認為 PM2.5 年平均濃度高於 10 μg/m3 或日平均濃度高於 25 μg/m3 時將影響人體健康。為有效監測這些數值，行政院環保署在全國設立了 84 個國家級監測站以評估大範圍的空氣品質，並定期維護校正以確保觀測數值正確性。另外，為加强區域性的空汙熱區監測，環保署另於城鄉增設價格較低廉的感測器，預計分年分區佈建 10,200 點形成「空氣品質感測網」，期待由綿密的感測器能更即時地監控小範圍的空汙數值，此感測網至 2021 年 6 月已佈建 7240 個測站。然而，由於區域感測器的儀器採用物理光學散射原理，容易受到環境因素影響造成較大的誤差，加上數量龐大故每台感測器的維護頻率低，單一區域感測器監控回傳的數值不如國家級測站來得可靠。當這些感測器出現異常時，並無法馬上判讀而維修，而造成許多應用這些資料的分析也出現誤差。而若需要維護這些感測器的品質，則需要長時間的監控感測器的數值是否有偏差，會耗費大量的人力以及時間資源，而且在目前的方法中，是採取隨機抽檢感測器，不能保證被檢測的感測器是否為真的有問題的感測器。因此，用現有的數據判讀感測器是否為故障感測器就極為重要，本篇論文就是希

望能深度探討這個問題。

1.2 方法簡介

為了降低感測器的維護成本並且增加感測器的可信度，本論文將使用台中的 466 個測站於 2018 年 1 月至 2018 年 12 月每分鐘的感測數據做為分析，將原始資料經過最後觀察值推估法 (Last Observation Carried Forward) 填補缺失值，並提出兩大類不同的時空資料整合方法，將填補過後的原始資料經過整合，最後輸入兩種不同的機器學習模型進行推估。由於我們的原始資料有包含由工研院提供的 2018 年隨機抽檢台中 146 個測站的結果，這是一個極為可貴的原始資料，能讓我們將其應用在半監督模型 (Semi-supervised)，於是我們將處理好的原始資料，匯入兩種不同的半監督模型: 第一個是先經過約束聚類 (constraint-based clustering)，再用我們所擁有的抽檢結果資料當作標籤 (label) 做主動學習，這個半監督模型稱為 SSDO (Semi-Supervised Detection of Outliers)。第二個是深度學習模型，稱為 Deep SAD。將這兩個不同的模型針對不同的輸入資料，進行了一系列的比較與分析，與原本的隨機抽檢以及經驗規則法則相比，則大大的提高了尋找異常感測器的機率。

1.3 論文貢獻

首先，我認為這篇論文能使大多數空汙相關的研究都能更為精準，因為能從最根本的地方解決數據可能存在不正確的問題，能盡量讓每次巡檢都能找出故障感測器並予以修復，讓這些較低廉而大量的感測器能更為可信。

第二，區域性空汙偵測器及市面上的廉價空汙偵測器大多以光學散射的原理來計算 PM2.5 的含量。然而，光學散射器容易因濕度、灰塵等變因造成偏差或故障，故需要耗費大量的人力成本進行測站的檢查及維修，目

前大多數單位採用的都還採用較為傳統的方式：包括隨機抽檢或依經驗派人至特定測站巡檢，本論文提出的方法將可以減少這方面的人力成本，因為在使用同樣的人力成本條件下，更能檢測到為故障的感測器，而在人力成本更高的國家效果將會更加明顯，因此，在台灣研究此議題我認為是非常適合的，由於地域較小，人口密度高，相鄰的感測器的距離相對起來也較小，能得到感測器標籤的巡檢也相對容易，因此本篇論文善加利用我們所擁有的巡檢結果，希望能在此議題的研究上有更大的幫助。最後，本論文所提出針對空污感測器的時空資料所結合的方法，並將此資料應用在半監督的模型中，不但保留了感測器的時空資訊，也加入了重要的標籤資料，並在最後的研究成果中有著很好的表現，而我們也討論了監督式模型、半監督式模型、無監督式模型在不同種空汙資料呈現型態的效果。

二、 相關研究

關於此 PM2.5 感測器異常偵測半監督模型的相關研究，我會分成兩個部份來介紹，一個是是關於 PM2.5 感測器異常偵測的部分，另一個是關於半監督異常偵測模型的相關研究。

2.1 PM2.5 感測器異常偵測相關研究

由於空氣汙染的存在一直都是大家很關心的議題，因此近年來也有許多關於空氣汙染異常偵測的論文，但此類問題大多是根據空氣汙染的資料，找尋可能有出現有不同於平時狀況的高汙染並將其稱為異常(Outlier)，大多都是關於此問題的異常偵測，而此類論文本質上與我們研究的問題不盡相同，但又有可參考之處，因此我們還是介紹幾篇關於此方向的論文。首先是由 V. M. van Zoest 等人提出的 Outlier Detection in Urban Air Quality Sensor Networks[1]，這篇論文是針對美國以及荷蘭 NO2 濃度資料，將觀測資料分為兩個空間和八個時間類別，提出了一種用於城市空氣質量傳感器網絡中離群值檢測的新方法，雖然與我們主要使用的機器學習模型的方法不同，是使用統計相關的方法找到離群值，但論文中也同樣提出時間與空間資料的重要，與我們的概念相同。再來介紹由 Fei Xiao 等人提出的 An improved deep learning model for predicting daily PM2.5 concentration[2]，這篇論文提出了加權的長期短期記憶神經網絡擴展模型（WLSTME），首先，選擇最近的周圍站點作為中心站點的相鄰站點，並將它們的距離以及空氣污染濃度和風況輸入到多層感知

（MLP）中，並將中心站點的歷史 PM2.5 濃度和相鄰站點的加權 PM2.5 系列數據輸入到長短期記憶（LSTM）中。與我們的論文之處是同樣選擇最近的周圍站點作為中心站點的相鄰站點，並且將時間與空間的資料做結合，而本篇論文所使用風向資料作為特徵，是值得我們參考集學習的。

接下來介紹的這篇論文是與我們論文主題非常接近，除了判斷空氣汙染中的異常 (Outlier)，也針對感測器的可信度進行推估計算，是由 Ling-Jyh Chen 等人提出的 ADF: An Anomaly Detection Framework for Large-Scale PM2.5 Sensing Systems[3]，此論文將空汙感測器的異常分為空間異常、時間異常、以及時空異常，並且針對這三個異常提出感測器故障檢測模組，並且針對不同的異常還能憑藉經驗給予不同的分析，像是裝置在室內的感測器、感測器裝設於污染排放源附近、故障的感測器、狀態無法判定的感測器。

這三篇論文同樣提出了時間與空間資料對於空氣汙染異常分析的重要性，因此本篇論文也著重於如何將空間與時間的資料作結合，而此類型的問題大多都缺乏了分析模型的可靠性，有就是標準答案的部分，本篇論文所應用了這三篇論文所未提及的標籤資料，因此我們更可以加以運用這些資料，將其分為訓練資料與測試資料，評估模型的可靠性，並且能使用我們接下來所述的半監督模型。

2.2 半監督模型異常偵測的相關研究

異常偵測是識別數據中異常樣本的任務，由於大多的問題都很難取得標籤，因此大多都是使用無監督學習的模型，像是由 Breunig, Markus M. 等人提出的 LOF: identifying density-based local outliers[4]，這是一種基於密度的異常值檢測，計算數據點周圍數據分佈的密度以發現異常，而除了此種基於密度的異常偵測外，也有人提出了截然不同的方法，像是由 Liu, Fei Tony 等人提出的 Isolation forest[5]，此方法使用了孤立二元數將異常資料區隔，是一個簡單而新穎的方法。而有些人認為這些淺層

的異常偵測模型有些限制，像是於到高維的特徵時需要有人工的特徵處理，以及在大型資料集中較難有擴展性，因此激發了這些人對深層的異常偵測模型的興趣，由 Ruff. 等人提出的 [6]，是一個深度神經網路的模型。

而在異常偵測的問題中，我們有時會擁有少量的標籤資料，因此，就有些人基於無監督的模型加以開發，將原本的無監督模型改為半監督模型，像本篇論文介紹由 Vincent Vercruyssen. 等人提出的 Semi-supervised Anomaly Detection with an Application to Water Analytics[7]，就是基於淺層無監督的方法加入了標籤系統，改成的半監督模型。而 Ruff. 等人也進一步開發了半監督的方法，Deep SAD: A Method for Deep Semi-Supervised Anomaly Detection[8]，就是一個深層的半監督模型。

三、 資料處理

我們所得到的資料包括 2018 年 1 月至 12 月台中的 468 個感測器以分鐘為單位的觀測資料，我們將刪減那些資料量過少的感測器，在我們篩選完之後剩下 466 個感測器，其中我們又擁有工研院在這一年巡檢的結果，也就是所謂的**標籤**，(詳見5.1.1) 因此我們將這些觀測資料經過一系列的處理，首先先填補缺失的資料，接下來使用特別的方法讓時空的資料做一個整合，最後再將資料擴展使我們擁有的資料量更大。這一系列的方法讓我們更能將 PM2.5 感測器的是否毀損的問題，應用在我們的半監督異常偵測模型中。

3.1 資料填補的方法

我們所用來分析的資料，需要 2018 年 1 月到 2018 年 12 月每個感測器每分鐘的值，若同一分鐘有超過一個以上的觀測值，則將該分鐘的所有觀測值取平均，將平均值當作該分鐘的觀測值。由於這種較為廉價的感測器並不會為了維護感測器的數據而不間斷的供電，所以會有長時間的或短時間的資料缺失，於是針對這兩種不同的缺失我們採用兩種不同的方法:

- 首先是短時間的資料缺失，若缺失資料少於 24 小時，我們會採用最後觀察值推估法 (Last Observation Carried Forward)，用缺失值前一個觀察值來填補該缺失值，短時間的資料缺失通常缺失的範圍不大，無論是使用平均值中位數或是簡單的插補方法等等做補值，

差異性並不大，因此我們採用最後觀察值推估法。

- 接著是長時間的資料缺失，由於空汙感測器具有時序性，所以若太長時間缺失資料，無法只使用前一個觀測值作為整段時間的觀測值，於是若缺失資料為連續且超過 24 小時以上，我們會使用前 24 小時的觀測值，來填補該缺失值。

3.2 將資料時空結合的方法

空汙感測器此類議題較為特殊的地方是在於觀測資料中無論是該感測器時序性的觀測資料，亦或是該感測器周圍感測器的空間資料，都扮演著非常重要的角色，於是該如何將空間及時間的資料整合，是此類問題一大探討的部分，於是我們提出了以下三種特別的整合方式。
我們將所要預測的感測器稱為中心感測器，而我們使用感測器的經緯度座標，計算出距離此感測器最近的五個感測器稱為周圍五個感測器，計算距離的方式如下[1]:

$$\mathrm{S} = 2\arcsin\sqrt{\mathrm{Sin}^2\frac{a}{2} + \cos(\mathrm{Lat\,1}) \times \cos(\ \mathrm{Lat2}\ ) \times \mathrm{Sin}^2\frac{b}{2}} \times 6378.137 \quad (3.1)$$

其中 Lung1 與 Lat1 表示第一個感測器的經度及緯度、而 Lung2 與 Lat2 表示另一個感測器的經度及緯度，而 a 為兩者的緯度之差，b 為兩者的經度之差，6378.137 為地球半徑，單位為公里。

我們的資料分布如下圖：橫軸為使用周圍 N 個測站當作鄰居，而縱軸為中心測站到周圍 N 個測站的距離，以公里為單位。我們以行政院環保署的空氣品質監測網中可以看到，智慧城鄉感測點、校園微型感測器以及

[1]資料來源：https://www.movable-type.co.uk/scripts/latlong.html

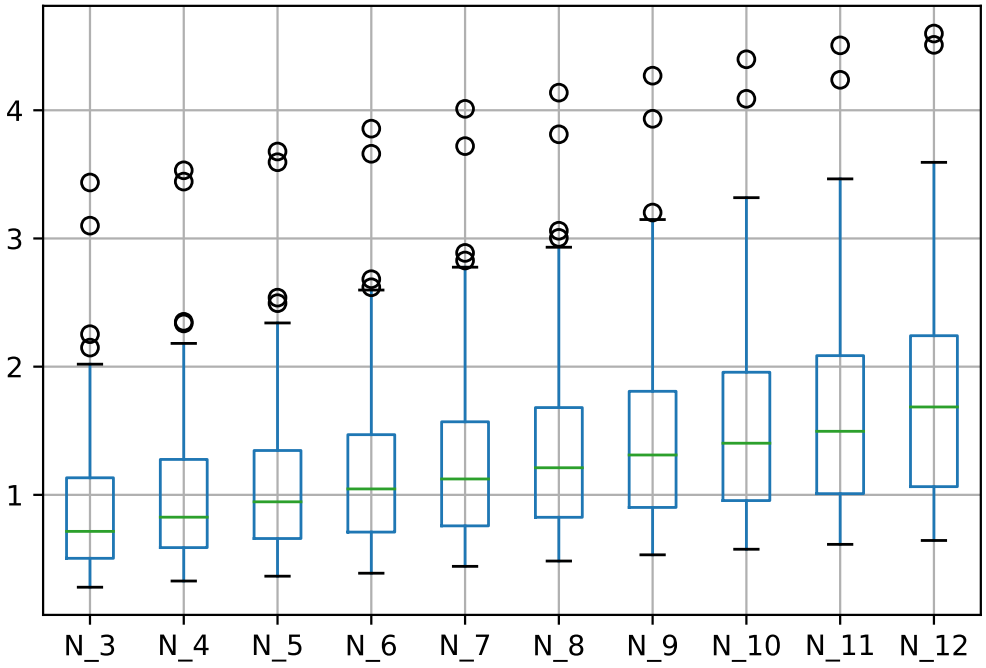

圖 3.1: 與周圍 N 個測站的平均距離

民間微型感測器的數據空間解析度約在一公里以內[2]，我們可以看到當鄰居為周圍五個測站時，平均的距離為一公里左右，因此我們取用周圍五個測站為鄰居，作為我們分析時使用的特徵資料。

3.2.1 使用圖片特徵整合時空結合的資料

有了中心測站及周圍的五個測站的資料後，我們想到了可以用圖片的方式整合時空的資料，於是我們用兩種不同的圖片類型來呈現時空結合的資料:

- 第一種是折線圖 (Line Graph)，橫軸為時間，從凌晨 0 點至晚上 12 點共 1440 分鐘; 綜軸為處理過後的 PM2.5 觀測值，而圖中我們使用黑色線來代表中心測站，並使用 5 條紅色線來呈現周圍的 5 個測站的觀測值，如此一來能在一張圖中清楚地掌握了時間的趨勢、以

[2]資料來源：空氣品質監測網 $https://airtw.epa.gov.tw/$

及周圍測站與中心測站的關係。如圖3.3和圖3.2。

- 第二種為熱力圖 (Heat Map)，相較起折線圖，熱力圖的資訊量較為精簡，橫向代表者中心感測器以及周圍五個感測器，總共會分成六個區域，縱向為一整天這些感測器的數值，由於我們的圖片使用28*28 像素，於是我們將 1440 分鐘分成 28 等分，所以縱向的一個區塊代表著每一等分平均下來的數值，並且使用不同的色階來表示。如圖3.4和圖3.5。

折線圖的優點是每一分鐘的數值，以及整天的時間趨勢完整的呈現出來;而熱力圖的優點則是將折線圖中較為繁雜的數字整合做平均，並且沒有像折線圖中有較多的留白部分，讓每一個像素都扮演重要的角色。

3.2.2 統整型資料

與上一個小節提出的圖片呈現方式相比，這個小節提出的方法更能展現出短時間內中心測站與周圍測站之間的關係，以及更為全面的資訊 [7]。由於在整合的數據中，若以天為單位則會讓資料難以呈現出較為精細的趨勢，於是我們將資料分成不同的區段 (window)，每一個區段為一個小時，但是在不同時段 PM2.5 值的變化非常大，所以我們使用的資料為各個測站在不同天中同一個小時的資料，以 PM2.5 濃度最濃的中午時段為例，我們會採用個個感測器中每一天的中午 12 點至 13 點的資料，而我們使用的資料分成下列三大類:

- 統計性特徵：中心測站在當個小時的最大值 (max)、最小值 (min)、平均值 (mean)、中位數 (median value)、標準差 (standard deviation)、偏度 (skewness)、峰態 (kurtosis)。總共為 7 個特徵 (Features)。

- 鄰近測站觀測差異之統計性特徵：除了中心測站的統計資料外，我們希望能有中心測站及周圍測站的關係，所以我們還會有中心測站

與周圍五個測站在當個小時每分鐘觀測值相減後的最大值 (max)、最小值 (min)、平均值 (mean)、中位數 (median value)、標準差 (standard deviation)、偏度 (skewness)、峰態 (kurtosis)。總共為 7 個特徵 (Features)。

- 描述性特徵：由於空汙資料中，時間的週期為非常重要的一個部分，無論是星期幾或是季節，都可能攸關著整個天氣的濃度變化，我們考慮兩個分類特徵：一周中的哪一天、一年中的季節為何，這些分類特徵使用一位有效編碼 (one-hot encoding) 處理，總共 11 個特徵。除了星期幾與季節之外，我們也發現時間，也就是一天中的幾點，是會很大影響空汙資料的一部分，目前我們是使用空汙濃度較濃的中午時段，而之後也可以將這一項當作描述性特徵使用，多 24 個特徵來表述是一天中的幾點。

3.2.3 統整型資料加上時序資料

為了不流失這一個區段中的時間序列資料，我們希望能取得這段時間內中心測站與周圍測站的時間序列分析，我們採用了 DTAI 團隊製作的時序動態扭曲 (DTW)[9] 演算法，來測量兩個時間序列的相似性。因此，我們將中心測站分別與周圍的五個測站這一小時的時序資料進行比對，取得共五個相似性分數，將這五個相似性分數也當為特徵來使用。

延承上一章所提出的概念，我們想在原有的資料中再加上中心測站的時序性資料，由於我們的每一個區塊是以小時為單位，於是我們除了3.2.2小節中提出的 25 個特徵，以及時序資料 5 個特徵之外增加了中心測站這 60 分鐘的觀測資料，將每分鐘當作一個特徵使用，因此，這一個小結提出的資料總共包括了 90 個特徵。

3.3 資料數量不足的解決方法

除了時空結合的問題之外，我們還面臨了資料量不足的問題，在第3小節中有提到我們只有 466 個感測器的觀測資料，因此若要拿來做深度學習模型的訓練資料稍嫌不足，於是我們想到了一種擴展資料的方式，我們使用一個月的資料，將資料以天為單位，分成 30 個不同的特徵向量，因此原來的資料筆數將擴展為 30 倍 (詳見5.1.1)，如此一來，我們就有較多的資料能放進我們的深度學習模型中使用。

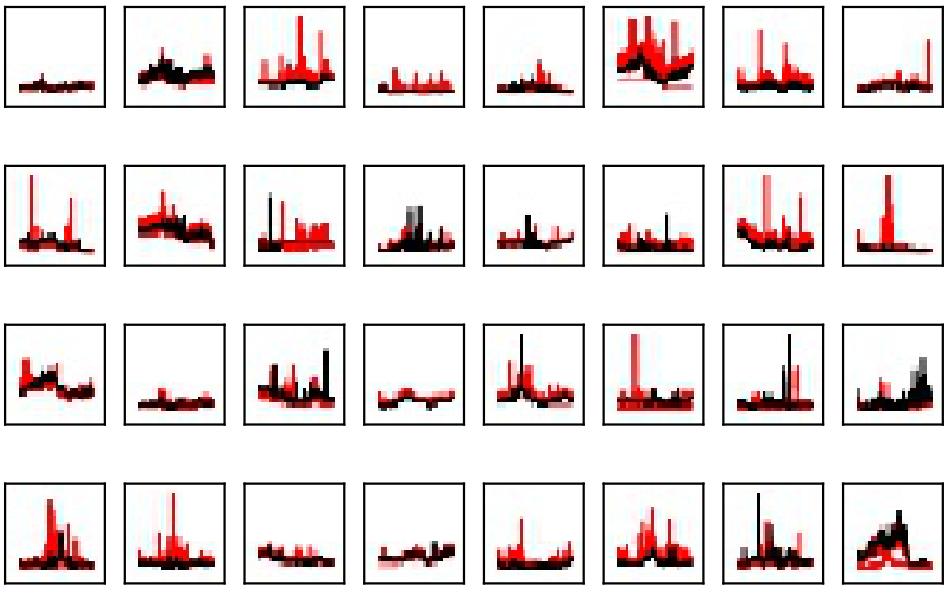

圖 3.2: 用折線圖表現的正常標籤資料

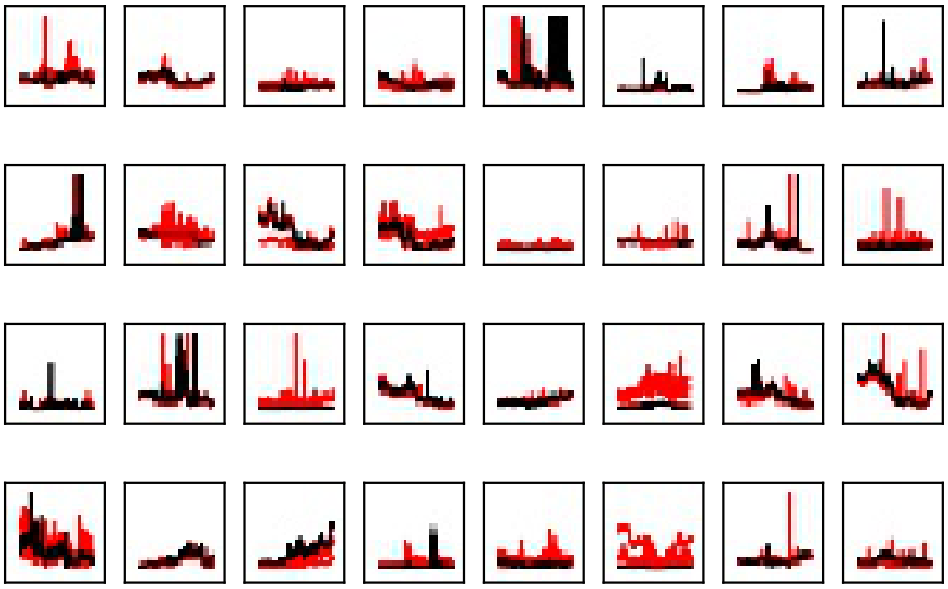

圖 3.3: 用折線圖表現的異常標籤資料

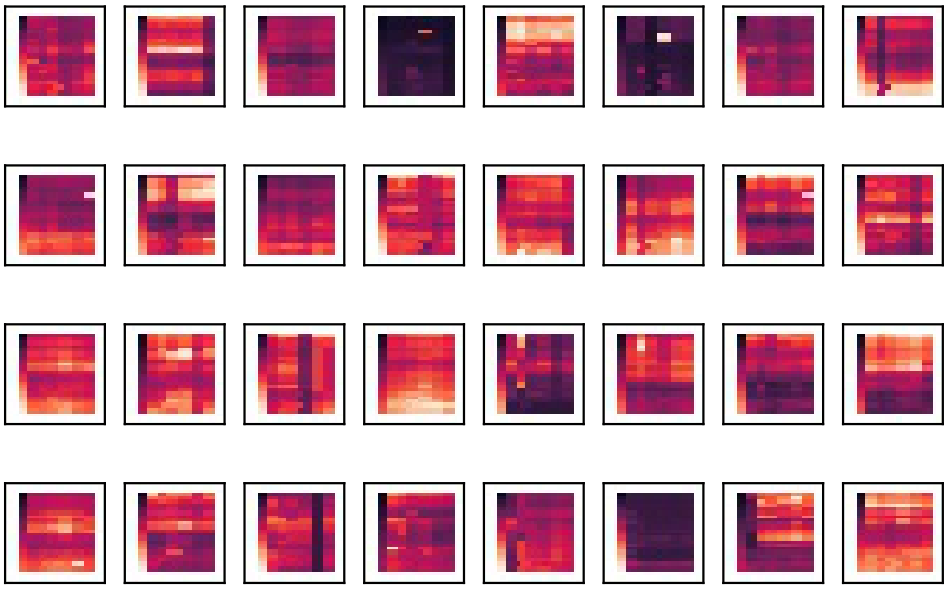

圖 3.4: 用熱力圖表現的正常標籤資料

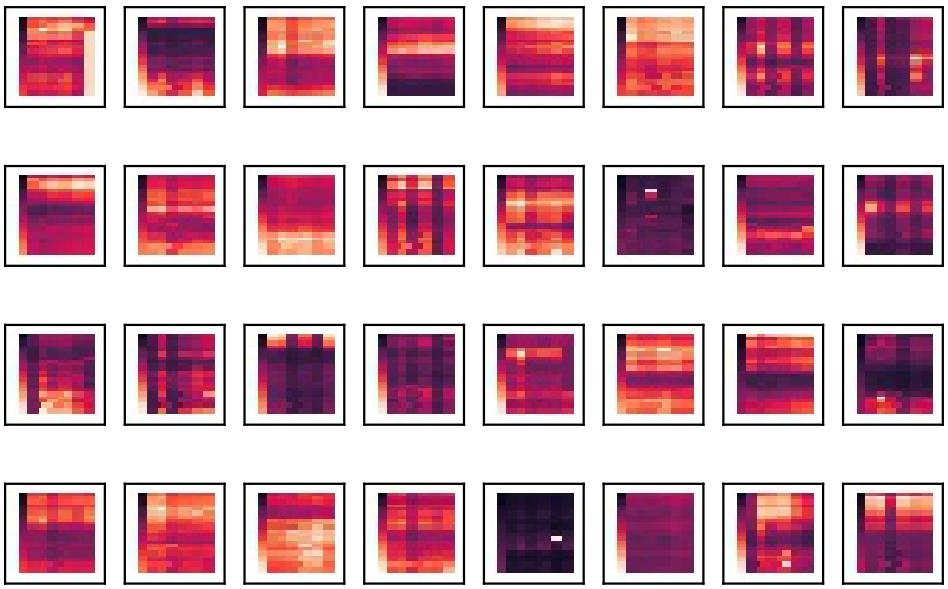

圖 3.5: 用熱力圖表現的異常標籤資料

四、 半監督模型介紹

上一個小節提出了三種不同的時空整合資料，而我們要將這些資料當成特徵資料，放進接下來要介紹的兩種半監督模型中。在空汙感測器是否異常的這類問題中，標籤 (labeled) 資料是非常非常重要的，因為巡檢的人力成本很高，並且需要長時間的量測，所以我們希望能最大化的應用這類標籤的資料，於是我們選用的兩種模型皆是半監督 (Semi-Supervised) 模型，希望能最大化的有效利用這些得來不易的標籤資料。接下來要介紹的兩種是不同類型的半監督模型，首先第一個 SSDO 是用聚類 (Clustering) 為基礎，再使用標籤進行標籤傳播與修正 [7]。第二個模型為 Deep SAD 是將無監督的異常偵測模型，修改為符合有少量標籤的半監督深度學習模型，希望在深層模型中，也能有很好的表現 [8]。

4.1 SSDO(Semi-Supervised Detection of Outliers)

SSDO 是一種新穎的、通用的方法 [7]，我們希望將這種基於約束聚類的異常檢測方法，應用於我們的空汙資料檢測中。SSDO 將應用數據 $F = \{f_0, \ldots, f_n\}$, 總共有 n 筆資料，其中 fi 代表著標準的特徵向量格式，會經由兩階段的過程給每個特徵一個異常分數。首先，會基於約束聚類發現的數據位置分布分配初始的異常分數。接著因為我們擁有標籤，所以會有標籤傳播階段，每個特徵向量的異常分數將會根據附近的已知標籤進行更新。

4.1.1 約束聚類 (Constrained Clustering)

SSDO 的第一步是對資料進行聚類，由於我們有標籤資料，所以我們可以用約束的形式指導聚類，我們再通過每一個正常和異常的特徵之間包含一個不可連接的約束 (cannot-link) 來完成約束聚類。我們不使用必須連接的約束 (must-link) 是因為不能保證所有正常或異常的行為都是相似的，例如: 可以有不同類型的異常行為，但是這些不同的異常行為不應該被迫出現在同一個群體中。我們採用 COP k-means 的演算法，若我們沒有標籤，也能使用標準的 K-means 演算法。在此聚類演算法中，異常特徵將看起來與正常的特徵不同，從幾何學上來說，異常特徵將遠離正常的特徵; 而從統計學的角度來看，意味著這些異常的特徵將位於空間中密度較低的區域。遵循這些假設，SSDO 這個方法定義出三個規則 (1) 一個特徵若離他的聚類中心 (cluster centroid) 越遠，則越有可能為異常。(2) 若一個特徵的聚類中心偏離其他的聚類中心越多，則此特徵可能為異常。(3) 若一個特徵所在的聚類越小，則此特徵可能為異常。基於這些假設，我們分別對每一個 $f \in F$ 分配以下的異常分數 $\text{score}(f)$:

$$score(f) = 1 - g\Big(\frac{point\_dev(f) \times cluster\_dev(f)}{cluster\_size(f)}; \gamma\Big) \tag{4.1}$$

接下來將針對其中的三個部分進行詳細的說明。

point deviation 將獲得 f 與聚類中心的偏差，將此群體的所有數據與中心的最大偏差進行標準化:

$$point\_dev(f) = \begin{cases} \frac{d(f,\ c(f))}{\max_{f_i \in C(f)} d(f_i,\ c(f))} & \text{if } |C(f)| > 1 \\ 1 & \text{otherwise} \end{cases} \tag{4.2}$$

其中 $C(f)$ 表示與 f 相同群體的所有特徵，$d(.)$ 表示歐幾里得 (Euclidean) 距離函數，$c(f)$ 是 f 所在聚類的聚類中心，如果 $C(f)$ 中只有一個特徵，則 point deviation 為 1。

cluster deviation 將獲得 f 的聚類中心有多不同於其他的聚類中心：

$$cluster\_dev(f) = \begin{cases} \frac{\min_{1 \le i \le n_c \wedge c(f) \neq c_i} d(c(f),\ c_i)}{\max_{1 \le i,j \le n_c} d(c_i, c_j)} & \text{if } n_c > 1 \\ 1 & \text{otherwise} \end{cases} \tag{4.3}$$

其中 c_i 表示聚類 i 的聚類中心，n_c 表示聚類的數量，如果只有一個聚類，則 cluster deviation 為 1。

cluster size 表示相對於擁有最大量特徵的聚類，聚類 f 所擁有特徵的數量：

$$cluster\_size(f) = \frac{|C(f)|}{\max_{1 \le i \le n_c} |C_i|} \tag{4.4}$$

其中，C_i 為被分配至聚類 i 的特徵集合。

squashing function 擠壓函數 g 將異常分數映射到 0 至 1 的範圍內，值越高表示異常程度越大：

$$g(x; \gamma) = 2^{-\frac{x^2}{\gamma^2}} \tag{4.5}$$

其中，參數 γ 將設置為使得映射後得分大於 0.5 的資料百分比。等同於用戶請求標籤時的標籤百分比，意思是說，這個模型是一個可以上線使用的模型，能及時的向用戶索取所需要的標籤，而參數 γ 就是最後要索取標籤的百分比。

4.1.2 透過已有的標籤進行更新分數

為了最大化具有標籤的價值，我們將上一小節所提到聚類分配的初始異常分數與標籤傳播獲得的分數相結合來更新它。為每個示例 f 獲得的更新分數為:

$$S(f) = \frac{1}{Z(f)} \left[score(f) + \alpha \sum_{f_j \in L_a} g\left(d\left(f, f_j\right); \eta\right) \right] \tag{4.6}$$

$$Z(f) = 1 + \alpha \sum_{f_j \in L} g\left(d\left(f, f_j\right); \eta\right) \tag{4.7}$$

其中，L_a 為標記異常樣本的集合，L 是所有標記樣本的集合。對於擠壓函數 (squashing function)$g(\cdot;\eta)$，η 是每個 $f' \in F$ 到他的 k^{th} 最近鄰居 (k-distance) 的距離的調和平均值，我們使用調和平均值是因為它對數據中的極值不太敏感，g 的目標是使標籤傳播的影響取決於標記和未標記特徵之間的距離，也就是說一個帶有標籤的特徵對附近特徵的得分比對遠處特徵的得分有著更大的影響。$Z(f)$ 是一個標準化的公式，使最後的分數在 0 至 1 之間，最後，α 控制標籤步驟相對於聚類步驟的影響力，若 α 大於 1，則只需要一個標籤則否決該聚類，若 α 較小，則需要多個標籤否決該聚類。

如同上述說到的，這是一個可以實時隨著索取到的標籤而做馬上調整的模型，當這個方法能線上應用時，可以更有效地得到所需要的標籤，例如異常分數在 0.5 左右的特徵，可以去實時的索取它的標籤，能讓這整個模型更完整，且更有效的利用了每一次的標籤。但目前尚未完成這一部份的實踐，可以是未來努力的一大方向。

4.2 Deep SAD(Deep Semi-supervised Anomaly Detection)

異常檢測大多被視為無監督的問題，而我們除了有大量的未標記樣本外，還有少部分由專家標記的標籤，因此我們將使用此異常檢測的半監督模型，除此之外，深度異常檢測方法最近在大型複雜數據集上顯示出優於淺層方法的有希望的結果，Deep SAD 就是一個應用於深度學習網路的半監督架構。為了明確我們的方向，首先我們會介紹深度無監督異常偵測模型 Deep SVDD(Deep Support Vector Data Description)[6]，然後將其推廣到半監督的異常偵測模型。

4.2.1 Unsupervised Deep SVDD

對於輸入空間 (input space) $\mathcal{X} \subseteq \mathbb{R}^D$ 與輸出空間 (output space) $\mathcal{Z} \subseteq \mathbb{R}^d$，令神經網路 $\phi(\cdot;\mathcal{W}) : \mathcal{X} \to \mathcal{Z}$ 是一個擁有 L 層隱藏層 (hidden layers) 和對應的權重 $\mathcal{W} = \{\boldsymbol{W}^1, \ldots, \boldsymbol{W}^L\}$。Deep SVDD 的目的是訓練一個神經網路 ϕ 去學習一種轉換，該轉換為最小化以輸出空間 $\mathcal{Z}$ 以點 $\boldsymbol{c}$ 為中心的封閉超球面體積。給定 n 個未標註的訓練資料 $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n \in \mathcal{X}$，Deep SVDD 的目標為:

$$\min_{\mathcal{W}} \quad \frac{1}{n}\sum_{i=1}^{n} \left\|\phi\left(\boldsymbol{x}_i;\mathcal{W}\right) - \boldsymbol{c}\right\|^2 + \frac{\lambda}{2}\sum_{\ell=1}^{L}\left\|\boldsymbol{W}^\ell\right\|_F^2, \quad \lambda > 0 \tag{4.8}$$

在 (4.8) 中，第一項為對樣本映射到超球面中心 $\boldsymbol{c}$ 的均方距離進行懲罰 (Penalize)，迫使網路提取那些最常見的變異子 (variation)，他們是資料集中最穩定的。因此，正常數據通常會映射到超球面的中心附近，而異常數據通常會映射得較遠 [6]。第二項資料為標準的權重衰減正則化 (weight decay regularizers)。

Deep SVDD 使用反向傳播 (backpropagation) 通過 SGD 進行優化，對於初始化，首先預訓練一個自動編碼器 (autoencoder)，然後用編碼器的收斂權重初始化網絡 ϕ 的權重 $\mathcal{W}$。初始化後，超球面中心 $\boldsymbol{c}$ 被設置為從前向傳遞中獲得的網絡輸出平均值。一旦網絡經過訓練，測試資料 $\boldsymbol{x}$ 的異常分數由神經網路的輸出 $\phi(\cdot;\mathcal{W})$ 到超球面中心的距離給出:

$$s(\boldsymbol{x}) = \|\phi(\boldsymbol{x};\mathcal{W}) - \boldsymbol{c}\|_2 \tag{4.9}$$

4.2.2 Deep SAD

接下來我們介紹半監督異常偵測模型 Deep SAD。我們現在擁有的資料除了 n 個未標註的訓練資料 $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n \in \mathcal{X}$ 且 $\mathcal{X} \subseteq \mathbb{R}^D$，我們還擁有 m 個有標註的資料 $(\tilde{\boldsymbol{x}}_1, \tilde{y}_1), \ldots, (\tilde{\boldsymbol{x}}_m, \tilde{y}_m) \in \mathcal{X} \times \mathcal{Y}$ 且 $\mathcal{Y} = \{-1, +1\}$，其中 $\tilde{y} = -1$ 代表著已知被標記為正常的資料，而 $\tilde{y} = +1$ 則代表著已知被標記為異常的資料。將 Deep SAD 網路的學習目標定為:

$$\min_{\mathcal{W}} \quad \frac{1}{n+m}\sum_{i=1}^{n}\|\phi(\boldsymbol{x}_i;\mathcal{W}) - \boldsymbol{c}\|^2 + \frac{\eta}{n+m}\sum_{j=1}^{m}\left(\|\phi(\tilde{\boldsymbol{x}}_j;\mathcal{W}) - \boldsymbol{c}\|^2\right)^{-\tilde{y}_j} + \frac{\lambda}{2}\sum_{\ell=1}^{L}\|\boldsymbol{W}^{\ell}\|^2 \tag{4.10}$$

Deep SAD 設定的學習目標中的未標記數據採用與 Deep SVDD 相同的損失項，因此若當我們沒有可用的標記數據 ($m = 0$) 的特殊情況下恢復為 Depp SVDD，也就是 [公式 (4.8)]。

對於有標籤的數據，Deep SAD 引入了一個新的損失項 (loss term)，為 [公式4.10] 的第二項，該損失項透過參數 $\eta > 0$ 來做加權調整，該參數控制著標記與未標記數據之間的平衡。當 $\eta > 1$ 時，更强調有標記的數據，而當 $\eta < 1$ 時，更强調未標記的數據。在 *Ruff.* 等人的實驗中，發現當 $\eta = 1$ 時在結果上有實質的提升，因此我們在此篇論文中的設定也將 η 設置為 1。

對於標記為正常 ($\tilde{y} = -1$) 的樣本，我們對映射點到中心點 $\boldsymbol{c}$ 的距離實施了平方損失函數，學習正常數據的潛在分布，而對於異常 ($\tilde{y} = +1$) 的樣本，我們將其懲罰了距離的倒數，使得異常樣本會映射得離中心點更遠。

五、 實驗結果

我們將上述提出的資料丟進兩個不同的半監督模型，而針對不同的資料型態在相同的模型中會有何影響，以及相同的資料型態在不同的模型中那些效果較好，並探討了給予模型正常及異常標籤量的多寡，以及未標籤資料的數量將會給模型帶來甚麼樣的影響。

5.1 資料介紹及實驗設置

5.1.1 資料介紹

就如上述所提到的，我們取得工研院關於 2018 年 1 月至 2018 年 12 月位於台中的 466 個感測器，其中有 146 個為工研院人工巡檢之後，有確認標註為正常或異常的感測器，並且包括 29 個標註為異常，而 117 個標註為正常。[表5.1] 但是巡檢的時間與人力有限，所以這些感測器的巡檢時間並不為同一天，是分散在 2018 年整年不同季節、不同月份的時間，所以我們取用資料時，無法採用一整年的資料當作訓練資料，於是我們採用的是巡檢日期前一整個月的資料當作我們所使用的資料，例如:A 感測器在 2018 年 5 月 2 日巡檢，則我們採用的資料 A 感測器及 A 感測器周圍感測器為 2018 年 4 月 2 日至 2018 年 5 月 1 日為止的資料。由於怕太過長時間的資料採取會影響我們的模型判斷，例如:B 感測器在 2018 年 12 月 1 日採檢為異常，但在 2018 年 1 月時可能還是正常的感測器，因此我們保守得只採用感測器前一個月的資料當作我們所使用的資

料。上述也有提到我們為了增加資料的使用，我們將這一個月的資料分成 30 天各別的資料，所以我們擁有的資料數量為有標註為異常資料的 29x30=870 筆、標註為正常資料的 117*30=3510 筆，以及未標註的感測器 320*30=9600 筆。

表 5.1: 工研院巡檢資料介紹

巡檢時間	巡檢感測器數量	正常	異常
2018-05	10	9	1
2018-06	35	28	7
2018-07	4	4	0
2018-09	60	49	11
2018-11	25	21	4
2018-12	12	6	6
總數	146	117	29

5.1.2 實驗設置

將上述的資料我們會將已標註的感測器依照 60/40 的比例分為訓練資料與測試資料，並將未標註的資料當作訓練資料的一部分，因此我們的訓練資料集包括了標註為正常的資料、標註為異常的資料以及未標註的資料，後面的章節也會提到關於這三者之間的多寡分別給予模型的影響。每個實驗都重複 10 次，使用不同的隨機訓練資料/測試資料分割，結果在 10 次實驗中取平均值。

5.1.3 比較的模型

由於目前討論此問題的論文大多都是未擁有標籤，大多都是開放式的提出解決方法並未有使用標準答案來驗證實驗結果的論文，所以這是一個十分新穎的題材，因此我們今天所用來比較的並不是其他學者所提出的論文結果，而是主要著重在討論各種不同模型及資料型態在此問題適配

度及實驗結果。

除了機器學習的模型之外，我們也比較了兩個非機器學習的方法，第一個是目前隨機巡檢的結果，也就是以表5.1來看，工研院在 2018 年隨機巡檢了 146 個感測器，其中有 29 個為異常，因此我們將此結果放入我們的比較模型中，而在第2節中有提到由 Ling-Jyh Chen. 等人提出的 ADF 方法，是使用根據經驗所得出的統計方式來判斷感測器的異常，而我們將此模型分成 ADF-5 及 ADF-10，前者代表著與我們的模型相同的設定，都只使用周遭的五個感測器進行判斷，後者代表著此模型設計得出最佳的結果，是以周遭的 10 個感測器為基準來進行統計。我們也與這個方法進行比較，看我們的模型是否有效。

而在機器學習的模型中，我們比較的方法有最基礎的監督式學習的回歸方式 (Linear regression、Ridge regression、Random Forest)[10][11][12]、以及無監督系統的孤立樹演算法 (Isolation Forest)[5]、深度模型的演算法包括了全監督的 MLP 神經網路 [13]，以及無監督的 SVDD，最後就是我們的方法 SSDO 及 Deep SAD，其中 SSDO 又分成兩種方式，第一種是上述所說使用有限制的聚類 (COP-kmeans) 當作無監督的分群方式，第二種則是使用孤立樹演算法 (Isolation Forest) 當作無監督的分群方式。

5.1.4 評量結果的方法

在我們的兩個模型中都會給予測試資料集的異常分數，將這個方法繪製成 ROC 圖，並且計算 ROC 圖曲線下面積為 ROCAUC，以這個為我們評量不同實驗結果的標準 [14]。除了 ROCAUC 之外，我們也針對了模型的精確度 (Precision) 和召回率 (Recall) 所計算出的 PR 曲線的曲線下面積 (PR-AUC) 進行比較 [15]。PR 曲線以召回率 (Recall) 為 X 軸，精確度 (Precision) 為 Y 軸，每一個點代表設定不同的門檻值 (Threshold) 所得到的不同的 Recall 及 Precision，最後繪製成一條曲線。而在 ADF

模型中我們將使用 Recall@k，以及 Precision@k，由於我們在此模型是使用原始資料，k 就相當於閥值，也就是當在 k 個感測器時，Precision 和 Recall 分別為多少。

5.1.5 超參數的設定

以下將分別介紹兩個模型的超參數設定:

SSDO：在此模型中重要的超參數有兩個，一個是在聚類階段要分成幾類的 n-cluster 我們稱之為 N，一個是在 [4.1.2] 中提到的，標籤傳播時將要影響最鄰近的 k 個特徵 (instance)，我們將此參數稱之為 K。

表 5.2: 調整超參數 N

超參數 N 與 K	ROCAUC
N=10，K=15	$\mathbf{0.7551 \pm 0.0040}$
N=15，K=15	0.7521 ± 0.0082
N=20，K=15	0.7462 ± 0.0124
N=25，K=15	0.7532 ± 0.0092

在表 [5.2] 中我們看到 N 參數在 {10,15,20,25} 中，以 N=10 表現最好，代表在 COP-kmeans 階段分群的數量太大反而會影響之後標籤傳遞的結果，所以在之後的實驗我們採用 N=10 的參數。

表 5.3: 調整超參數 K

超參數 K 與 N	使用的無監督模型	ROCAUC
K=10，N=10	SSDO with iforest	0.7676 ± 0.0120
	SSDO with COP-kmeans	0.7620 ± 0.0092
K=15，N=10	SSDO with iforest	$\mathbf{0.7780 \pm 0.0044}$
	SSDO with COP-kmeans	0.7551 ± 0.0040
K=25，N=10	SSDO with iforest	0.7647 ± 0.0033
	SSDO with COP-kmeans	0.7503 ± 0.0065
K=35，N=10	SSDO with iforest	0.7654 ± 0.0074
	SSDO with COP-kmeans	0.7510 ± 0.0107

在表 [5.3] 中我們看到 K 參數在兩種不同的無監督分群方式中，有著不同的結果，K 在 {10、15、25、30} 中，當 K=10 時，使用 COP-kmeans 的方法有較好的結果; 而當 K=15 時，使用 iforest 的方法有更好的結果，而其中又以使用 iforest 的 SSDO 表現得更好，於是我們之後的實驗採用了 K=15，但兩者的共同點是當 K 越大有略微下降或是趨於不穩定的趨勢，代表著加入標籤後影響周圍的特徵數量不能太大，否則會失去標籤的作用。

Deep SAD：在此模型中較為重要的超參數是批量 (batch size)，我們使用了三種不同的批量作為參考。而公式4.10中的參數 η ，如同我們所說的，在 *Ruff.* 等人的實驗中，發現當 $\eta = 1$ 時在結果上有實質的提升，因此我們在此篇論文中的設定也將 η 設置為 1。

表 5.4: 調整 Deep SAD 的 batch size

	ROCAUC
batch size=64	0.9009 ± 0.0060
batch size=128	$\mathbf{0.9028 \pm 0.0054}$
batch size=256	0.8965 ± 0.8151

在表 [5.4] 中我們看到當 batch size 為 128 時有較好的結果，因此之後的實驗我們將 batch size 設置為 128。

5.2 實驗結果與討論

在本章節我們主要會針對以下三個大問題進行實驗及討論:

* 問題一：不同的模型在空汙感測器異常的實驗結果
* 問題二：不同的資料型態能在何種模型中展現出最佳的結果
* 問題三：若調整輸入進模型中的異常標籤、正常標籤、未標註資料的數量將會對此空汙異常偵測有何種影響

5.2.1 不同的模型的比較及實驗結果的探討

我們將上述所提到的四種不同的資料型態，有折線圖、熱力圖、廣泛統整型資料、以及統整型資料加上時序資料 (綜合資料)，每一種轉換成一種特徵向量，輸入到我們在第 [5.1.3] 節中所提出的各式模型中，而 ADF 方法與我們的資料型態不相符，因此我們使用原始資料來做實驗。想探討出何種模型在帶有少量標註的空汙感測器的異常偵測實驗中有最好的結果。在這個實驗中，我們所擁有 m 筆已標記，以及 n 筆無標記的訓練資料，而我們的模型包刮了以下三種:

1. 監督式學習 (Supervised learning): 用 m 筆資料當作訓練資料。
2. 半監督式學習 (Semi-Supervised learning) 用 $m+n$ 筆做訓練資料。
3. 無監督式學習 (Unsupervised learning) 當作未擁有標記資料，用 $m+n$ 筆資料筆做訓練資料。

而我們的 SSDO 模型也能夠當作監督式模型來學習，也就是只使用 m 筆資料當作訓練資料，當作未擁有無標籤的資料。因此有了以下這個實驗結果:

在表 [5.5] 中我們可以看到綜合所有實驗結果，Deep SAD 在使用 MLP 網路架構的 ROCAUC 能到 0.9 以上，這代表著我們無論是資料的呈現方式以及模型的選用都有著非常好的表現，而在 SSDO 中，ROCAUC 也有 0.8 左右的結果，比起無監督孤立樹 (Isolation Forest) 演算法的 0.59，代表標籤傳播這一塊扮演著不可或缺的重要性。而監督式的回歸模型最高也來到了 0.7 左右，讓我們更確定了這種理論，標籤在空汙感測器的異常偵測實驗中是非常重要的。

我們能從表 [5.6] 中看出工研院一年中隨機抽檢 146 個感測器中，只有 29 個為異常的感測器，大約為 0.2 的比例能找出有問題的感測器，若能使用我們所提出的方法先選出異常分數較高的感測器做採檢，會有較高的機率找出有問題的感測器，讓巡檢的效益達到最高。

表 5.5: 不同模型在不同資料中所得到的 ROC-AUC

ADF-5		0.6240 ± 0.0000		
ADF-10		0.6940 ± 0.0000		
	折線圖	熱力圖	統整性資料	統整及時序資料
Linear regression	0.6138 ± 0.0124	0.6433 ± 0.0126	0.6034 ± 0.0140	0.5938 ± 0.0082
Ridge regression	0.6414 ± 0.0124	0.6924 ± 0.0071	0.6035 ± 0.0140	0.5938 ± 0.0082
Random Forest	0.6138 ± 0.0124	0.6433 ± 0.0126	0.6034 ± 0.0140	0.5938 ± 0.0082
MLP supervised learning	0.5523 ± 0.0118	0.5785 ± 0.0121	0.6386 ± 0.0067	0.6507 ± 0.0112
SSDO with iforest	0.7912 ± 0.0041	0.8118 ± 0.0036	$\mathbf{0.7961 \pm 0.0028}$	$\mathbf{0.8034 \pm 0.2394}$
SSDO with COP-kmeans	$\mathbf{0.7913 \pm 0.0048}$	$\mathbf{0.8128 \pm 0.0039}$	0.7890 ± 0.0106	0.7891 ± 0.0091
Isolation forest	0.4958 ± 0.0140	0.5132 ± 0.0107	0.5433 ± 0.0129	0.5924 ± 0.0075
Deep SVDD	0.5703 ± 0.0083	0.6434 ± 0.0109	0.5766 ± 0.0078	0.5793 ± 0.0078
SSDO with iforest	0.7849 ± 0.0036	0.7760 ± 0.0052	0.7369 ± 0.0105	$\mathbf{0.7730 \pm 0.0029}$
SSDO with COP-kmeans	0.7892 ± 0.0038	0.7554 ± 0.0112	$\mathbf{0.7529 \pm 0.0058}$	0.7621 ± 0.0079
Deep SAD	$\mathbf{0.8989 \pm 0.0069}$	$\mathbf{0.9028 \pm 0.0549}$	0.6441 ± 0.0129	0.6823 ± 0.0125

表 5.6: 不同模型在不同資料中所得到的 PR-AUC

隨機巡檢			0.1940 ± 0.0000	
ADF-5			0.2900 ± 0.0000	
ADF-10			0.4400 ± 0.0000	
	折線圖	熱力圖	統整性資料	統整及時序資料
linear regression	0.2769	0.3137	0.3339	0.3163
ridge regression	0.3214	0.3876	0.3337	0.3159
random forest	0.3290	0.4292	**0.4471**	**0.4588**
SSDO with iforest	0.3374	0.4555	0.3061	0.2883
SSDO with COP-kmeans	0.3399	0.5158	0.3177	0.2554
Isolation fores	0.1886	0.2003	0.2375	0.2578
SSDO with iforest	0.3712	0.4114	0.2645	0.3773
SSDO with COP-kmeans	0.3640	0.4162	0.2809	0.3214
Deep SAD	**0.8099**	**0.8048**	0.3450	0.4215

5.2.2 整合時空的資料型態探討

在表 [5.5] 中能看到這四種資料型態在此實驗中都有很好的結果，代表我們在處理此種時空整合的空汙資料的方法有達到一定的效果。而其中折線圖及熱力圖在 Deep SAD 的表現是最好的，而廣泛的統整型資料在 SSDO 中也好於它在 Deep SAD 中的表現，所以我們推斷每一種資料呈現的型態適合的模型都不相同，之後的研究方向也許可以著重探討不同的資料型態所分別擁有的優點是甚麼，以及適合使用的模型為何，若能將這兩種的優點整合，那又能使我們的模型更上一層樓。

5.2.3 調整給予模型標記爲正常、異常及未標記的比例

我們前面有說到，給予模型有標籤的特徵資料是一件很重要的事，但是當我們有一定數量的標籤資料後，該給模型多少具有標籤的資料，又或者是該如何分配標籤資料中正常及異常的比例，以及這些標籤分別給予模型的影響力等，都是十分值得探討的問題，因此我們這個小節就是做了一系列的實驗，希望能讓這些標籤發揮最大的效用。本章節針對此問題設計了三種實驗：若我們所擁有 m 筆已標記，以及 n 筆無標記的訓練資料

1. 我們將固定 n 筆資料，並且調整 m 筆資料中標記為正常及異常的資料比例。此實驗想探討給予模型標籤資料的數量及正常異常的比例會帶給模型何種影響。若對應到現實的情況，就是當我們巡檢完一定數量個感測器之後，該給予模型多少比例的正常及異常的資料。

2. 我們將固定住 m 筆資料，並且調整 n 筆資料量的大小，看整體會有何變化。此實驗想探此實驗想探討當標籤數量固定時，未標記的數量是否越多越好。若對應到現實的情況就是當我們只有能力固定巡檢一定數量的感測器，是否給予模型那些未能巡檢的感測器資料，並且要給予多少。

3. 我們將固定住 $m+n$ 筆資料，並且將 n 逐漸加大及 m 逐漸減小，此實驗想探討若感測器數量固定，是否越多的標籤資料效果越好。若對應到現實的情況，就是當我們擁有的感測器數量固定，若給予模型不同的已巡檢數量，是否會讓模型有更好的表現。

5.2.3.1 調整給予模型標籤的數量及正常異常的比例

而我們將 SSDO 及 Deep SAD 分開討論，SSDO 我們使用的資料型態是3.2.3中所提到的統整型資料加上時序資料，而 Deep SAD 使用3.2.1中所提到的熱力圖作為模型所使用的資料。

SSDO: 在這個模型中，一開始的聚類階段，使用的 COP-kmeans 演算法，就會將標籤為正常及標籤為異常的資料給予不可連接 (cannot-link)，而在後面的標籤傳播階段，這兩者也大大影響著周遭的異常分數，因此到底該給予模型多少標籤的資料以及兩者的比例對這個實驗來說非常重要。

從第 [5.1.1] 節中的資料介紹中能看到我們擁有 3510 筆標記為正常的資料，以及 870 筆標記為異常的資料，而我們依照 60/40 的比例分為訓練資料與測試資料，也就是 2160 筆標記為正常的訓練資料及 522 筆標記為異常的訓練資料。我們先固定標記為異常的資料為全部的 0.02，也就是 10 筆資料，並調整標記為正常的資料從全部 0.004 遞增為 1，也就是從 10 筆遞增到 2160 筆。接下來我們再將標記為異常的比例調高至 0.2，也就是 100 筆，而標記為正常的資料如同上述的方法遞增。接下來是將標記為正常的資料分別固定為 0.004 及 0.04，也就是 10 筆及 100 筆，而調整異常的比例從 0.02 遞增至 1，也就是從 10 筆遞增至 522 筆。我們將標籤為正常的調整比例 normal-labeled-ratio 簡稱為 n-ratio，而標籤為異常的調整比例 anomaly-labeled-ratio 簡稱為 a-ratio。

表 5.7: SSDO 模型中固定參數 a-ratio，調整參數 n-ratio

	使用的無監督模型	a-ratio=0.02	a-ratio=0.2
n-ratio=0.004	SSDO with iforest	$\mathbf{0.7676 \pm 0.0072}$	0.7367 ± 0.0116
	SSDO with COP-kmeans	0.7551 ± 0.0040	0.7350 ± 0.0143
n-ratio=0.04	SSDO with iforest	0.7571 ± 0.0032	0.7595 ± 0.0080
	SSDO with COP-kmeans	0.7565 ± 0.0056	0.7560 ± 0.0080
n-ratio=0.4	SSDO with iforest	0.7561 ± 0.0038	0.7586 ± 0.0042
	SSDO with COP-kmeans	0.7536 ± 0.0060	0.7601 ± 0.0076
n-ratio=1	SSDO with iforest	0.7518 ± 0.0044	0.7571 ± 0.0046
	SSDO with COP-kmeans	0.7533 ± 0.0074	0.7532 ± 0.0070

表 5.8: SSDO 模型中固定參數 n-ratio，調整參數 a-ratio

	使用的無監督模型	n-ratio=0.004	n-ratio=0.04
a-ratio=0.02	SSDO with iforest	0.7367 ± 0.0116	0.7571 ± 0.0032
	SSDO with COP-kmeans	0.7350 ± 0.014	0.7565 ± 0.0056
a-ratio=0.2	SSDO with iforest	0.7350 ± 0.012	0.7595 ± 0.0080
	SSDO with COP-kmeans	0.7346 ± 0.014	0.7560 ± 0.0080
a-ratio=0.55	SSDO with iforest	0.7172 ± 0.0083	$\mathbf{0.7730 \pm 0.0029}$
	SSDO with COP-kmeans	0.7228 ± 0.010	0.7621 ± 0.0079
a-ratio=1	SSDO with iforest	0.7157 ± 0.0027	0.7684 ± 0.0052
	SSDO with COP-kmeans	0.7243 ± 0.0073	0.7561 ± 0.0047

從表 [5.7] 中我們可以看到，當標記為異常的資料給的較少時，發現就算增加正常標記的比例，也未必會給模型起到多大的作用，同理，在表 [5.8] 中當正常資料給的少時，就算異常資料給的再多也一樣沒有好轉，只有讓標記為正常及異常的資料同時增加，並且有較好的比例時，才會讓這

個模型發揮最大的效用。

Deep SAD: 在這個實驗中，我們先固定正常資料的比例為 0.5，這裡的 0.5 是指在所有有標籤的資料中佔有多少的比例，有標籤的訓練資料包括異常與正常共計 2628 筆, 這裡的 0.5 就是指 1314 筆。並調整異常的比例從 0.01 開始遞增至 0.15, 這裡的 0.01 及 0.15 同樣是指在所有有標籤的資料中佔有多少的比例, 我們將標籤為正常的調整比例 normal-labeled-ratio 簡稱為 N-ratio，而標籤為異常的調整比例 anomaly-labeled-ratio 簡稱為 A-ratio。

表 5.9: Deep SAD 模型中調整參數 N-ratio，固定 A-ratio=0.15

	ROCAUC
N-ratio=0.0	0.9086 ± 0.0086
N-ratio=0.1	$\mathbf{0.9100 \pm 0.0089}$
N-ratio=0.5	$\mathbf{0.9100 \pm 0.0089}$

表 5.10: Deep SAD 模型中調整參數 A-ratio，固定 N-ratio=0.5

	ROCAUC
A-ratio=0.01	0.6926 ± 0.0210
A-ratio=0.05	0.8075 ± 0.0117
A-ratio=0.10	0.8692 ± 0.0100
A-ratio=0.15	$\mathbf{0.9100 \pm 0.0089}$

在表 [5.9] 中看到，當我們增加標註為正常的資料標籤給模型時，並不會有顯著的成長，在 Deep SAD 的模型中，標記為正常的資料並不是主要影響模型學習的關鍵，反之，我們看到表 [5.10] 中，當標記為異常的資料

提高時，就能夠讓模型的準確率有非常大幅度的提升，因此又更一步確定了我們的推論，標籤在異常偵測模型中有著非常大的影響，而其中，標記為異常的標籤又更為重要。

5.2.3.2 調整給予模型未標籤的數量

在我們的模型中，除了有標籤的資料外，佔大部分的其實是未標記的資料，於是我們想說若之後我們擁有的有標籤的資料量來到了一定的數量，是否能減少未標記的資料讓我們模型更為穩定呢，這個章節就是在探討未標記的資料多寡對模型的影響。我們同樣把 SSDO 與 Deep SAD 分開討論，與上述情況相同，SSDO 我們使用的資料型態為3.2.3中所提到的統整型資料加上時序資料，而 Deep SAD 使用3.2.1中所提到的熱力圖作為模型所使用的資料。

SSDO: 在這個實驗中，從表 [5.1] 我們原先擁有的未標記數量為 9660 筆，我們將陸續減少供應給模型的未標記數量，而我們將此比例稱為 Unlabeled-ratio，簡稱為 u-ratio。

表 5.11: SSDO 模型中調整參數 u-ratio

	使用的無監督模型	ROCAUC
u-ratio=1	SSDO with iforest	0.7730 ± 0.0029
	SSDO with COP-kmeans	0.7621 ± 0.0079
u-ratio=0.6	SSDO with iforest	0.7715 ± 0.0036
	SSDO with COP-kmeans	0.7612 ± 0.0149
u-ratio=0.3	SSDO with iforest	0.7850 ± 0.0033
	SSDO with COP-kmeans	0.7694 ± 0.0052
u-ratio=0	SSDO with iforest	$\mathbf{0.8034 \pm 0.0023}$
	SSDO with COP-kmeans	0.7891 ± 0.0091

從表 [5.11] 中可以明顯的看到，當未標註的資料放入越少時，模型的準確

率會上升，我們推測是因為在我們的實驗設置中，無監督的分群會影響最後實驗的結果，而我們因為資料量不足，所以將少量的無監督資料以天數分割為較大量的資料，而其中若有存在實際上為異常的感測器，會因為此原因增加此資料的數量。造成分群時有可能因為此群體資料量大而未被察覺為異常，因此當我們的數據量足夠時，就不需要此種以天數分割資料的方式，我相信這會讓我們的模型有更好的表現。

Deep SAD: 在這個實驗中，我們想實驗在不同的給予模型異常標籤數量下 (為5.2.2中 A-ratio)，不同比例的未標籤資料會如何影響模型，因此我們會固定兩種不同的 A-ratio，並餵入模型遞減的未標籤資料，將此比例稱為 U-ratio。

表 5.12: Deep SAD 模型中調整參數 U-ratio

	A-ratio=0.1	A-ratio=0.15
U-ratio=1.0	0.8692 ± 0.0100	0.9100 ± 0.0089
U-ratio=0.8	0.8761 ± 0.0068	0.9111 ± 0.0102
U-ratio=0.6	0.8803 ± 0.0107	0.9133 ± 0.0064
U-ratio=0.4	0.8854 ± 0.0055	0.9167 ± 0.0069
U-ratio=0.2	0.8892 ± 0.0052	0.9199 ± 0.0056
U-ratio=0.0	0.8904 ± 0.0096	$\mathbf{0.9200 \pm 0.0052}$

在這個實驗中與上一個實驗展示出的結果相同，將模型中的未標籤資料減少，反而會讓模型的效果變好，我們推測是因為在 Deep SAD 中，因為假設異常資料會遠小於正常資料，但由於我們上述提到的增加資料的方法，導致未標籤資料中的異常資料與正常資料以相同的倍數增加，因此我們在未標籤資料越多的情況訓練效果越差，有可能是這個原因。改善方法為我們在上述提到的，若能夠得到足夠多的未標籤資料，則不需

要以這種增加資料數量的方法訓練模型，能讓真正的標籤正常與異常的資料、以及未標籤資料在模型中有最好的發揮。

5.2.3.3 感測器數量固定，是否越多的標籤資料效果越好

在這個實驗設置中使用 Deep SAD 模型，我們將訓練資料固定為 2628 筆，並將未標記的資料從全部的 0.95 降至 0.05，也就是從 2496 筆降至 131 筆，我們稱之為 un-ratio，而標記的資料從 0.05 增加之 0.95，我們稱之為 la-ratio。其中，標記為正常與異常的比例從頭到尾都保持相同比例 80/20。我們希望將資料筆數固定，查看是否其中越多標籤資料對結果越有幫助。

表 5.13: 固定資料筆數，調整標籤與未標籤資料的比例

un-ratio=0.95，la-ratio=0.05	0.6740 ± 0.0055
un-ratio=0.75，la-ratio=0.25	0.7541 ± 0.0019
un-ratio=0.50，la-ratio=0.50	0.8908 ± 0.0032
un-ratio=0.25，la-ratio=0.75	0.9191 ± 0.0045
un-ratio=0.05，la-ratio=0.95	$\mathbf{0.9412 \pm 0.0010}$

由 [表5.13] 可以清楚的看到，當我們擁有較多的標籤資料時，模型的效果有明顯的增長，與我們上述所得到的結果相同，標籤資料是很重要的，也可以解讀成，在有限資源中，我們將感測器的數量固定，若給予模型越多有標籤的感測器，模型效果會越好。

5.2.4 是否給予預訓練的影響

在 Deep SAD 的模型中，我們將自動編碼 (autoencoder) 的預訓練 (pre-training) 用於初始化。該自動編碼具有與公式4.10網路 ϕ 相同架構的編碼器，經過訓練後，我們使用該編碼器收斂的初始化參數 $\boldsymbol{W}$。

表 5.14: 預訓練的影響

Pre-train or not	Deep SAD ROCAUC
yes	$\mathbf{0.9028 \pm 0.0054}$
no	0.8929 ± 0.0072

這個實驗的設置為 A-ratio=0.15，N-ration=0.5，U-ratio 為 1，我們得到有預訓練的效果比沒有預訓練的結果更好，因此我們在 Deep SAD 的實驗都會使用自動編碼器的預訓練。

六、 總結

6.1 結論

本篇論文針對 PM2.5 空汙感測器，提出了圖片特徵、統整形資料、統整時序型資料，這三種將時空資料結合的方法，善用了很珍貴的空汙感測器巡檢結果當作標籤資料，將上述的資料應用於深層及淺層兩種不同的半監督異常偵測模型中。而我們發現不同的資料呈現型態適合的模型可能有些不同，像是圖片型資料適合的可能是深度的半監督模型，而統整形資料及統整時序型資料適合淺層的半監督模型。

為了證實模型的效用，我們將 SSDO 模型以及 Deep SAD 模型與現在使用的隨機巡檢方法、淺層的無監督及全監督模型、以及深層的無監督及全監督模型進行比較，我們使用的半監督模型的效果優於無監督模型，代表標籤傳播這一塊扮演著不可或缺的重要性。而我們的半監督模型也優於全監督的回歸模型，因此我們認為使用未標記的資料調整模型，有一定的效果。而比起隨機巡檢方法，我們的模型有大幅度的提升，證實我們提出的資料結合方法以及應用於半監督模型都是非常有效果的，能夠有較大的機率找到故障的感測器。

除此之外我們也討論了標籤資料量以及未標籤資料量兩者間的調整以及正常與異常標籤帶給模型的影響，我們發現擁有較多的標籤資料時，模型的效果有明顯的增長，而其中標記為異常的資料又更為重要，而標

記為異常的資料越多，反而會讓模型的效果下降，我們推測是因為我們的為資料量不足，所以擴展資料的方法增加了原本可能存在的標記為異常的感測器，造成模型的表現下降，因此當我們的數據量足夠時，就不需要此種以天數分割資料的方式，在未來有更大量資料時能有更好的調整結果。

6.2 未來展望

首先，在資料方面，第3章資料處理中有提到，我們因為資料量不足，因此用了擴展天數的方法來强制增加資料量，若之後我們能擁有更大量的巡檢資料，以及更多感測器的觀測資料，相信能在模型的效果上有更好的提升。而關於其中的統整形資料，我們發現時間也是重要的一環，之後也可以將時間的部分擴展為特徵資料。

接著，在第2章相關研究中有提到的資料種類的研究，因為我們現在所用的資料只有純粹感測器的 PM2.5 值，若之後能用專業知識融合風向、風速、濕度、温度等等的資料，也能給模型更多的特徵加以學習。

而在模型改善的部分，我們有提到 SSDO 可以是一個可以線上應用主動學習模型，當我們的模型得出某些特別不確定的感測器是否有問題時，若能及時巡檢並回報，相信又會是另一個里程碑。我們也想到若能將 CNN 等捲積模型用於我們的圖片資料上，也會是另外一種可以嘗試的方式。

參考文獻

[1] V. Van Zoest, A. Stein, and G. Hoek, "Outlier detection in urban air quality sensor networks," *Water, Air, & Soil Pollution*, vol. 229, no. 4, pp. 1–13, 2018.

[2] F. Xiao, M. Yang, H. Fan, G. Fan, and M. A. Al-Qaness, "An improved deep learning model for predicting daily pm2. 5 concentration," *Scientific Reports*, vol. 10, no. 1, pp. 1–11, 2020.

[3] L.-J. Chen, Y.-H. Ho, H.-H. Hsieh, S.-T. Huang, H.-C. Lee, and S. Mahajan, "Adf: An anomaly detection framework for large-scale pm2. 5 sensing systems," *IEEE Internet of Things Journal*, vol. 5, no. 2, pp. 559–570, 2017.

[4] M. M. Breunig, H.-P. Kriegel, R. T. Ng, and J. Sander, "Lof: Identifying density-based local outliers," in *Proceedings of the 2000 ACM SIGMOD international conference on Management of data*, 2000, pp. 93–104.

[5] F. T. Liu, K. M. Ting, and Z.-H. Zhou, "Isolation forest," in *2008 eighth ieee international conference on data mining*, IEEE, 2008, pp. 413–422.

[6] L. Ruff, R. Vandermeulen, N. Goernitz, L. Deecke, S. A. Siddiqui, A. Binder, E. Müller, and M. Kloft, "Deep one-class classification," in *International conference on machine learning*, PMLR, 2018, pp. 4393–4402.

[7] V. Vercruyssen, W. Meert, G. Verbruggen, K. Maes, R. Baumer, and J. Davis, "Semi-supervised anomaly detection with an application to water analytics," in *2018 ieee international conference on data mining (icdm)*, IEEE, vol. 2018, 2018, pp. 527–536.

[8] L. Ruff, R. A. Vandermeulen, N. Görnitz, A. Binder, E. Müller, K.-R. Müller, and M. Kloft, "Deep semi-supervised anomaly detection," *arXiv preprint arXiv:1906.02694*, 2019.

[9] W. Meert, K. Hendrickx, and T. V. Craenendonck, *Wannesm/dtaidistance v2.0.0*, version v2.0.0, Aug. 2020. DOI: 10.5281/zenodo.3981067. [Online]. Available: https://doi.org/10.5281/zenodo.3981067.

[10] G. A. Seber and A. J. Lee, *Linear regression analysis*. John Wiley & Sons, 2012, vol. 329.

[11] A. E. Hoerl and R. W. Kennard, "Ridge regression: Biased estimation for nonorthogonal problems," *Technometrics*, vol. 12, no. 1, pp. 55–67, 1970.

[12] A. Liaw, M. Wiener, *et al.*, "Classification and regression by randomforest," *R news*, vol. 2, no. 3, pp. 18–22, 2002.

[13] F. Rosenblatt, "Principles of neurodynamics. perceptrons and the theory of brain mechanisms," Cornell Aeronautical Lab Inc Buffalo NY, Tech. Rep., 1961.

[14] J. A. Hanley and B. J. McNeil, "The meaning and use of the area under a receiver operating characteristic (roc) curve.," *Radiology*, vol. 143, no. 1, pp. 29–36, 1982.

[15] J. Davis and M. Goadrich, "The relationship between precision-recall and roc curves," in *Proceedings of the 23rd international conference on Machine learning*, 2006, pp. 233–240.