腸病毒預報：應用機器學習預測腸病毒疫情

Jiun-Yi Tsai 蔡均宜
資訊工程學系
國立中央大學
桃園，台灣
honey1209@gmail.com

Jia-Ying Shih 史家瑩
資訊工程學系
國立中央大學
桃園，台灣
okok861120@gmail.com

Hung-Hsuan Chen 陳弘軒
資訊工程學系
國立中央大學
桃園，台灣
hhchen@g.ncu.edu.tw

***Abstract*—在本文中，我們探討了容易增加腸病毒感染率的因素，採用政府公開資訊，以線性迴歸模型、隨機森林、支援向量機、及 XGBoost 套件所實作的梯度提升技術等模型來預測未來一週台北及桃園地區腸病毒的疫情狀況，表現最好的模型 R^2 分數約為 0.9，顯示我們可以有效透過機器學習模型預測腸病毒疫情。我們目前已將預測模型的原始碼開源，並計劃將預測結果製成腸病毒預報，透過網路平台和通訊軟體發布，針對腸病毒的預防，提供社會大眾作為參考，從而減少腸病毒感染人數。**

Keywords – Enterovirus, Predictive models, Diseases, Machine Learning, Linear Regression, Random Forest, Support Vector Machine, XGBoost

1. 導論

腸病毒有很強的傳染性，特別是每年的 5 月、6 月、及 9 月是腸病毒的好發期。腸病毒可能經由糞口、食物、飛沫、或病人的分泌物等污染途徑向外擴散。每個人感染腸病毒後症狀的嚴重程度各有不同，有不小的比例的感染者只會有輕微的症狀，而使得患者本人可能會沒有意識到自己以感染腸病毒，或誤會自己感染的是其他病毒而掉以輕心。若無症狀的患者在群體之中，很容易因為衛生習慣差或是互動密切等因素，導致群體中的其他人也遭受傳染，最後造成群聚感染。在所有類型的腸病毒中，又以腸病毒 71 型最為可怕，腸病毒 71 型不僅有可能會破壞心臟、腦部等重要器官，進而引發嚴重併發症，甚至可能導致死亡 [1]。腸病毒感染症目前並沒有特效藥，因此事先做好預防措施以防止腸病毒傳播是很重要的。

本論文應用機器學習預測腸病毒疫情，並計劃提供預測結果於網路平台上，讓社會大眾作為防疫參考。我們探討了容易增加腸病毒感染率的因素，得知 5 歲以下孩童特別容易感染腸病毒，且腸病毒感染與氣溫、空氣品質等大氣環境有關。因此採用縣市的腸病毒門診人次、縣市相對溼度、縣市 0-5 歲人口資料、縣市氣溫、縣市 PM2.5 濃度等五組資料作為訓練模型所需的數據，我們嘗試了四種不同的監督式模型以預測未來一週台北及桃園的腸病毒案例數，並以 R^2 分數及方均根分數 (root-mean-square error) 作為預測模型與結果是否相符的判別標準。

我們嘗試了線性迴歸模型 (Linear Regression)、隨機森林 (Random Forest)、支援向量機 (Support Vector Machine)、及 XGBoost 套件中所實作的梯度提升技術 (Gradient Boosting) 做為預測模型。結果顯示：以 XGBoost 的梯度梯升技術作為預測模型時，能得到最好的效果，最終訓練出的模型 R^2 分數在桃園市及台北市的結果分別為 0.90 及 0.92，代表模型預測的數值與實際數值十分接近。

我們計劃將預測數據視覺化，透過圖表呈現腸病毒疫情趨勢，使其更易於理解，並透過網路平台發布腸病毒疫情預報，讓民眾能及時獲得預報訊息，提前採取防疫措施，防範腸病毒感染，進而趨緩整體疫情。

本文將檢閱過去傳染病數量預測的相關研究及腸病毒的相關環境變因 (Section 2)；說明我們採用的預測模型及使用的變因 (Section 3)；回報模型的準確性 (Section 4)；說明我們正在開發中的預警平台 (Section 5)；最後並討論此作品的可能限制及未來的方向 (Section 6)。

2. 相關研究

本節回顧過去流行性傳染病的數量預測及預警平台的相關研究，以及腸病毒流行的相關環境變因。

A. 流行性傳染病數量預測

流行性傳染病的傳播與疾病本身的特性及人群間的接觸有關。最知名的流行性傳染病數量預測模型可能是隔間模型 (Compartmental model) [2]，此模型列出得病至痊癒的各個狀態 (例如：可受感染 Susceptible、已感染 Infectious、已復原 Recovered 等)，並以各狀態間的轉移機率 (transition probability) 預測每個狀態隨時間變化的人數，進而預測最終的得病人數。然而，這種模型僅能估計總體的得病人數與健康人數，無法更細緻地考慮人與人的接觸而評估每個人的染病機率或傳染途徑等。

近期許多作品從「社群網路」的角度切入，試圖能從個體的層次推測傳染病的傳播途徑與每個個體的染病機率。這類型的作品通常需要假設個體間的傳播條件 [3]，進而推斷最後疾病傳播的廣度，簡易且被廣泛採用個體間傳播條件模型包括線性閥值模型 (Linear Threshold model) [4] 與獨立傳播模型 (Independent Cascade model) [5] 等。以這些傳播方式為基礎，科學家們也研究該如何有效地中斷傳播路徑 (例如：該優先對哪些人施打疫苗、該優先隔離哪些人) 來讓減少染病人數 [6]。

然而，在實務上我們很難得知兩兩個體間的實體接觸歷史，因此要建構出實體接觸的「社群網路」窒礙難行。但同時，我們也明瞭實體接觸對於腸病毒傳播的重要，最後我們採取折衷的作法：由於地理位置相近的群體比較有機會接觸，故為每一個行政單位單獨建構預測模型，這個方法不用建構接觸網路但又兼顧一部份的區域特性，應該是實務上較可行的做法。

B. 腸病毒的環境因素與感染人數

腸病毒雖然是透過接觸而傳播，但研究顯示環境因素會影響到腸病毒的活躍程度，這些環境因素包括：氣溫、濕度、空氣品質等。例如：氣溫與腸病毒感染人數在統計上有顯著的關係 [7], [8]，雨量與腸病毒感染人數也呈現正相關 [8]。

以上這些文章雖然展示了可能的環境因子，但並未採用機器學習的模型預估腸病毒的染病人數，本文則利用這些可能的因子做為一部份的特徵來預測未來一週的腸病毒的染病人數。

C. 傳染性疾病預警平台

就我們所知，國內目前較知名的傳染性疾病預警管道有兩個：第一個是「流感預報站」[1]網站，該網站使用每週流感門診歷史數據和機器學習模型來預測未來四週的流感疫情，並將預測結果圖像化後，每週定期更新於網站上供民眾瀏覽。此網站的資料呈現方式清晰，我們考慮在網站上採用類似的資料呈現方式。然而，流感預報站並未提供使用者主動訂閱的服務，故關心此議題的大眾需要自行回訪此網站以取得新資訊，使用上比較不方便。另外，此網站對於預測模型的細節並未說明，亦未公開原始碼，因此比較難將成果擴散採用至其他傳染性疾病。第二個較有名的平台是衛生福利部疾病管制署的「疾管家」[2]Line bot 聊天機器人，這個管道的好處是平台能主動推播訊息至訂閱者的手機中，故使用者只要被動地接受訊息即可。但目前此平台公佈的資訊是「已發生」的事件說明，並未提供未來的預測功能，也未提供歷史趨勢圖表。流感預報站及疾管家是兩個獨立的服務，我們則計劃同時開發腸病毒的預報網站及 Line bot，兩者可互相連結。

3. 確診數預測模型

本節說明訓練資料的取得及資料前處理與確診數預測模型的訓練方式。

A. 預測模型與特徵選擇

我們使用了四種常見的監督式學習模型來預測下一週的腸病毒就診人數，包括：線性迴歸模型、隨機森林、支援向量機、及梯度提升技術，其中，梯度提升技術我們採用 `XGBoost` 套件 [9]，這是最近幾年的資料科學競賽中贏家常用的一個梯度提升技術套件，與標準梯度提升技術最大的不同在於：(1) `XGBoost` 加上正則化 (regularization) 以防止過適；(2) `XGBoost` 採用損失函數的二階導數以加速參數更新。

[1]https://fluforecast.cdc.gov.tw/

[2]https://page.line.me/vqv2007o

我們參考了多篇與腸病毒有關的論文後，發現腸病毒的活躍程度和環境因素有很大的關聯性，例如：腸病毒感染人數和氣溫呈現正相關 [7], [8]，也和累積降雨量正相關 [8]。除此之外，亦有研究顯示：PM2.5 體積小，會直接入侵肺泡，此時免疫系統的巨噬細胞會吞噬 PM2.5，但由於 PM2.5 是無機物無法被消滅，就好比吃下石頭，反而造成人體的免疫系統癱瘓，使身體更難抵禦其他病毒的攻擊 [10]。因此，我們選擇了這些環境量測值做為模型的一部份特徵。

另一方面，0 至 5 歲的兒童為腸病毒的好發族群，再加上幼兒免疫力較成人差，感染腸病毒的患者多為 5 歲以下的孩童。因此我們也在模型中加入了 0 至 5 歲人口數作為訓練的特徵。

綜合以上所述，我們最後選擇了以下五種特徵：0 歲至 5 歲兒童人口數、溫度、相對濕度、PM2.5 濃度、及週次。我們蒐集自 2008 年第 14 週起到 2020 年第 34 週為止，共 646 週上述特徵的資料，以這些特徵來預測接下來一週的腸病毒就診人數。在資料前處理上，由於我們使用的訓練特徵都是數值資訊，且少有遺漏，於是我們只將缺失的特徵值補入該特徵的平均值。我們分別自政府開放資料庫[3]下載「健保門診及住院就診人次統計 – 腸病毒」資料集、「急診傳染病監測統計 – 腸病毒」資料集，從內政部戶政司全球資訊網[4]取得了各縣市的單齡人口資料，在環保署的環境資源資料庫[5]下載空氣品質監測值，從中提取氣溫、PM2.5 濃度、相對溼度等資料，再統整成所需的資料格式。

在資料的切割上，我們最後選擇將 70% 的資料 (共 455 筆) 當作訓練集，並將另外 30% 的資料 (共 135 筆) 做為測試集。每個模型我們都花了一定的工夫嘗試重要的超參數的組合，讓每個模型儘可能發揮較大的效果。

4. 實驗結果

本節敘述各個預測模型的預測準度 (以 R^2 分數及 RMSE 分數呈現之)，並分析各種特徵的重要性。

A. 預測模型的準度比較

表 I
各模型的 R^2 分數與 RMSE 比較表 (桃園)

模型	R^2 分數	RMSE
線性迴歸	0.18	619.46
隨機森林	0.87	250.55
支援向量機	0.76	335.22
`XGBoost`	0.90	214.59

我們首先選擇了我們所在的城市 – 桃園市，作為疫情預測的目標地區。表 I. 展示了實驗結果。除了線性迴歸模型表現得比較普通，其餘的幾個模型都表現得還不錯。其中，使用 `XGBoost` 建出來的模型 R^2 分數為四個模型中最高的，約為 0.90。

我們選擇的第二個城市，是人口眾多且密度高的台北市。一樣是以週次、台北市的 0-5 歲兒童人口數、溫度、

[3]https://data.gov.tw/

[4]https://www.ris.gov.tw/app/portal

[5]https://erdb.epa.gov.tw/

表 II
各模型的 R^2 分數與 RMSE 比較表（台北）

模型	R^2 分數	RMSE
線性迴歸	0.17	529.28
隨機森林	0.88	202.88
支援向量機	0.86	219.54
XGBoost	0.92	159.26

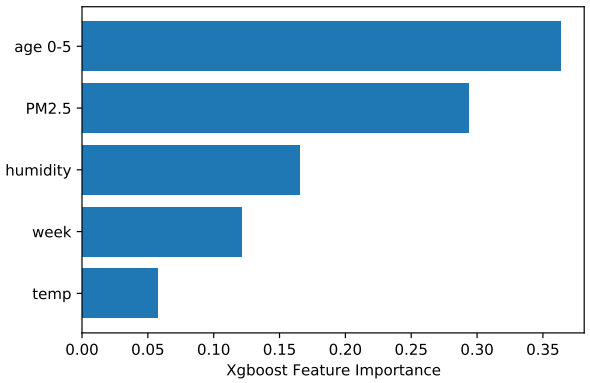

圖 1. 各個特徵的重要性

相對濕度、及 PM2.5 濃度作為特徵訓練模型，預測台北市每週可能的腸病毒就診人數。從表 II 中可見結果和表 I 的結果吻合：採用 XGBoost 建出來的模型表現最好，R^2 分數約為 0.92；而表現最糟的模型仍是線性迴歸，R^2 分數僅有 0.17。

B. 特徵值的重要性比較

由於 XGBoost 的預測效果最好，我們藉由此套件的 XGBRegressor 所訓練出的模型中的屬性 feature_importances，幫助我們比較各個特徵的重要性，其中 feature_importances_ 的計算方式是在每棵樹依據不純度 (impurity) 對特徵做排序後，再取整個森林的平均，不純度的值越大代表該特徵的重要性越高。圖 1 呈現的特徵重要性。

從圖 1 中，我們可以得知模型中各個特徵的重要程度，其中以 0-5 歲的人口數影響程度最大，PM2.5 指數次之。相對溼度、週數、及溫度的影響則比較小。這裡的週數指的是以年為單位的第幾週，若假設一月一日至一月七日為第一週，則一月八日至一月十四日為第二週，以此類推。加入週次是因為腸病毒是一種與季節息息相關的傳染病。舉例來說，腸病毒的感染人數通常都會在每年的 5、6 月達到高峰後，又逐漸趨緩。

5. 預警平台規畫

圖 2 展示預警平台的設計架構，我們將透過網路爬蟲，每週從各資料集的來源網站抓取新的資料存入資料庫，並持續訓練模型產生新的預報。

訓練出腸病毒預報模型後，我們預計架設腸病毒預報網站，網站的界面設計如圖 3 所示。我們以彩色長方色塊將使用者導向各個縣市的預測結果頁面，點擊後將進

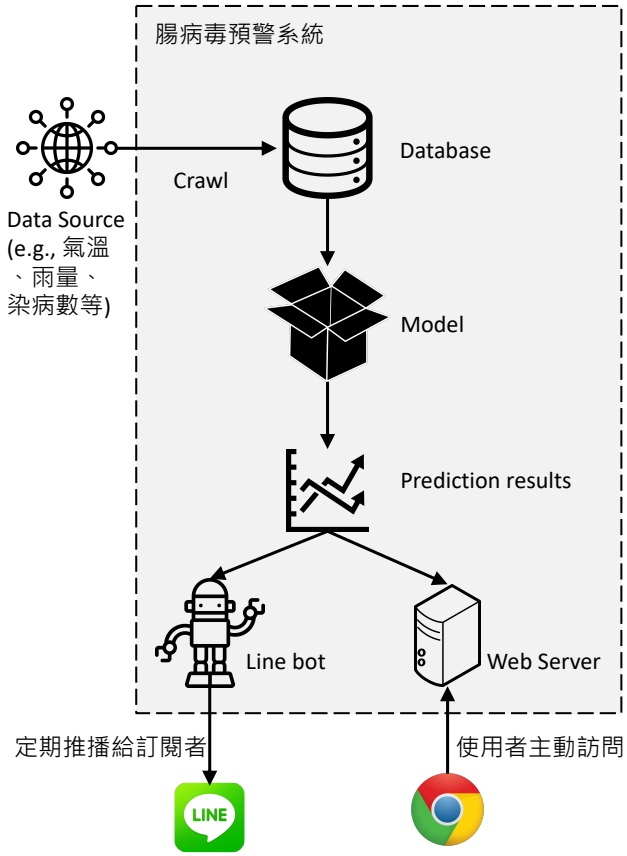

圖 2. 資料串接設計架構圖

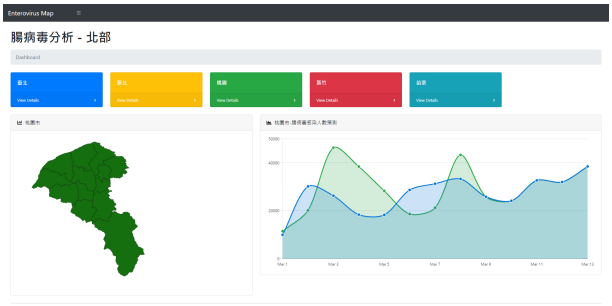

圖 3. 網站 UI 設計圖

一步顯示該縣市的腸病毒地圖與就診人數預測趨勢折線圖。為使預測數據視覺化讓預報更易於理解，我們會將各區域疫情變化與預測趨勢繪製成折線圖表 (如圖 4)，其中綠色折線代表實際就診人數，而藍色則為模型預測的就診人數趨勢。而我們也會依據疾管署訂定的腸病毒流行閾值繪製「腸病毒地圖」，把地圖中腸病毒案例超過該地區流行閾值的區域以不同顏色標示，並將以上資訊每週更新、發布於腸病毒預報網站。

除了網站，我們也預計透過 LINE bot 發布腸病毒預報通知，在有較高風險時即時發送警示訊息給訂閱者。預報內容包含「所在地區的感染人數」、「疫情發展趨

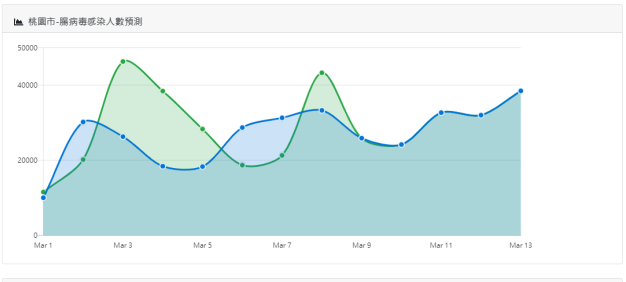

圖 4. 疫情預測趨勢表

勢」、「高風險區域」等，並定期每一週到兩週推播一次衛生教育的訊息提醒使用者。

6. 討論與未來展望

幼兒感染腸病毒可能引發嚴重的呼吸道或神經系統症狀及併發症，且預估在台灣每年五歲以下幼兒因腸病毒產生的醫療成本超過三千萬元 [11]，若能有效預測腸病毒的未來趨勢並及早預警，應可大幅降低幼兒感染率及照護成本。我們的研究發現：使用簡單的機器學習模型搭配政府公開資料即可有效預測未來一週各縣市的幼兒腸病毒感染人數，若將預測結果搭配上網站及聊天機器人，應可做為有效的預警平台。我們計劃針對 0 歲至 5 歲幼兒的照顧者 (如家長及幼兒園老師等) 進行宣傳，目前我們已經與中央大學附設幼兒園的園長及老師聯繫，等平台正式上線後將以他們為第一波的推廣對象，讓他們能夠在模型預測到腸病毒可能即將發生大流行時，提高警覺，注意身邊孩童的衛生習慣，降低腸病毒的感染率。

我們已公開預測模型的原始碼，[6]讓其他研究者能複製我們的實驗；同時，我們也會儘可能公開網站的視覺化模組和聊天機器人的推播模組的原始碼，未來若有其他流行性傳染病有類似的需求，將能透過我們公開的程式快速複製此成果。

我們正在進行以及近期內預計開工的有以下幾個部份。第一，目前我們尚未完全自動化從政府公開資料平台爬取資料的過程。若此預警系統要正式上線，我們需要讓系統每天能自動截取公開平台上的資料，並在整份資料有缺漏時自動發出通知告知系統管理員，或者在資料僅有部份特徵值有缺失時自動進行合理的補值。第二，我們目前僅針對台北市及桃園市訓練模型並進行預測，平台上線時我們預計至少應包括六都 (台北市、新北市、桃園市、台中市、台南市、及高雄市) 的預測資訊。

最後，倘若本預警系統吸引足夠多的使用者，當系統預計下週為高風險週次，且使用者真的因為高危險預警而做出應對措施 (如：勤洗手、避免出入公共場所等)，這些應對措施可能因而減低腸病毒的染病人數，讓預警模型的預測「看似」不準確。再說得宏觀一些，凡是使用者會因預測而改變行為，進而又因為不同的行為而影響到結果的動態環境，我們似乎難以確認預測究竟準確或不準確。我們覺得這是一個值得深入研究的有趣問題，但目前我們仍不確定該從何下手。

[6]https://github.com/lelebilu/Enterovirus-epidemic-forecast

References

[1] K. Y. Lee, "Enterovirus 71 infection and neurological complications," *Korean journal of pediatrics*, vol. 59, no. 10, p. 395, 2016.

[2] W. O. Kermack and A. G. McKendrick, "A contribution to the mathematical theory of epidemics," *Proceedings of the royal society of london. Series A, Containing papers of a mathematical and physical character*, vol. 115, no. 772, pp. 700–721, 1927.

[3] H.-H. Chen, Y.-B. Ciou, and S.-D. Lin, "Information propagation game: a tool to acquire humanplaying data for multiplayer influence maximization on social networks," in *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2012, pp. 1524–1527.

[4] W. Chen, Y. Yuan, and L. Zhang, "Scalable influence maximization in social networks under the linear threshold model," in *2010 IEEE international conference on data mining*. IEEE, 2010, pp. 88–97.

[5] K. Saito, R. Nakano, and M. Kimura, "Prediction of information diffusion probabilities for independent cascade model," in *International conference on knowledge-based and intelligent information and engineering systems*. Springer, 2008, pp. 67–75.

[6] J. Leskovec, A. Krause, C. Guestrin, C. Faloutsos, J. VanBriesen, and N. Glance, "Cost-effective outbreak detection in networks," in *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2007, pp. 420–429.

[7] 莊慈容, "台灣腸病毒感染之季節變動與相關環境因子探討," 2011.

[8] 陳國良, "評估氣候變遷對腸病毒重症之經濟影響-以台灣地區為例," 2011.

[9] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," in *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, 2016, pp. 785–794.

[10] B.-J. Lee, B. Kim, and K. Lee, "Air pollution exposure and cardiovascular disease," *Toxicological research*, vol. 30, no. 2, pp. 71–75, 2014.

[11] D.-P. Liu, T.-A. Wang, W.-T. Huang, L.-Y. Chang, E.-T. Wang, S.-H. Cheng, and M.-C. Yang, "Disease burden of enterovirus infection in taiwan: Implications for vaccination policy," *Vaccine*, vol. 34, no. 7, pp. 974–980, 2016.