國立交通大學資訊工程學系
資訊專題成果報告
以機器學習預測氣喘病患未來氣喘狀況之App
An app for predicting asthma patient's future condition with machine learning

專題題目說明、價值與貢獻自評：

此專題以氣喘病患平時紀錄之各項生理資料，搭配天氣空污等數據，導入機器學習模型，並製作成 App 讓病患平時可紀錄各項生理數據，展示未來氣喘嚴重程度情形，並幫助其監控及預防氣喘之發生。

專題隊員：

學號	姓名	手機	E-mail	負責項目說明	專題內貢獻度(%)
0613457	陳韋霖	0928899896	willyc.cs06@nctu.edu.tw	資料前處理、模型調參 App server 端程式、報告撰寫	45
0616027	陳昱銘	0984077391	tim310579.cs06@nctu.edu.tw	資料前處理、特徵工程 App client 端程式、報告撰寫	45

本專題如有下列情況則請說明：
1.為累積之成果(含論文及專利)、2.有研究生參與提供成果、3.為大型研究之一部份。

黃宇負責提供各項機器學習模型調整之建議，也幫忙部分報告的修改，對於實驗規劃也給予許多良好的指導。

相關研究生資料（無則免填）：

級別年級	姓名	提供之貢獻	專題內貢獻度(%)
博士班	黃宇	實驗及報告建議	10

【說明】上述二表格之專題內貢獻度累計需等於 100%。

指導教授簡述及簡評：

韋霖和昱銘在進行本專題的過程十分認真，並積極主動的解決問題與改進，專題成果相當優異。
未來本專題成果預期可朝收集 app 使用者心得及更完整的醫療資料繼續做優化和發展，對智慧醫療照護領域具有高度發展潛力。

指導教授簽名：曾新穆

中華民國一〇九年 十二月 二十三日

目錄

一、摘要

氣喘是一種因體質或外在因素刺激導致的慢性呼吸道疾病，發作可能緩慢也可能非常快速，且氣喘目前是兒童最常見的慢性疾病，急性發作危機常常是兒童和家長害怕、恐懼與不確定感的來源，故監控病情十分重要。另外台灣的氣喘病患人數也有增加的趨勢，根據台灣地區健保資料庫的資料顯示，從 1995-1996 年約有 3.7% ~ 7.1%的青少年曾被診斷為氣喘，2000-2007 年，20 歲以下兒童及青少年被診斷為氣喘的比例為 15.7%，到了現在，台灣更是有 20%的小孩有氣喘，可見氣喘在台灣已成為一項不可忽視的疾病，因此本專題希望能以預防的角度，除了病患本身的資料以外，再加入環境的因素，幫助氣喘患者及早得知氣喘的發作可能，以便早一步做好防範。

本專題透過病患紀錄之生理資料，搭配天氣、空污等相關資料，預測病患未來氣喘之發作嚴重程度情形。經實驗結果分析，對於較嚴重病情之召回率(recall)達 80%，可準確且有效的預測嚴重的氣喘情況。

二、簡介

本專題使用氣喘病患平時紀錄的生理資料，配合日期加上天氣和空污資訊作為輸入，建立多種機器學習之模型相互比較，並從中選出實驗結果較佳之模型，用於預測未來氣喘情形。

製作 App，整合模型，方便日常使用與紀錄，並能幫助平時病情的監控。本專題的 App 旨在達成及早得知該氣喘病患隔日之氣喘情形，並告知使用者以做早一步之預防和準備。

三、問題描述

因氣喘發作之不確定性，導致氣喘病患即使有定期回診治療及用藥，但在天氣或環境劇烈變動時，還是會擔心自身何時發作，但其實氣喘之發作情形在日常生活中都是有跡可循，病患平時除了記錄自身狀況外，較少會注意天氣、空污等會影響氣喘發作之相關資訊，導致氣喘發作之時機不好掌控及預防，本專題就是以此為出發點，在病患平時有做生理數據紀錄的同時，再加入環境的因素，給予其一些預防的建議。

四、文獻與現有系統調查比較

氣喘是指受到過敏原或刺激物刺激後，支氣管產生慢性發炎的情形，依其嚴重的程度可以呈現呼吸困難、喘鳴音、胸悶和咳嗽等症狀，嚴重影響個人生活品質。雖然大部分的氣喘目前無法治癒，但是可以藉由良好的控制來減少或預防氣喘發作，Julie L. Harvey[5]和 Sathish A. P. Kumar[5]就曾在論文中提到，如何使用機器學習之模型，及早發現氣喘孩童並給予治療，而我們的專題則是在病患已有氣喘的情況下，預測短期氣喘情形的變化並加以預防。

關於預測短期氣喘的部分，陳佳妤[6]在他的論文中是使用決策樹（decision tree）及資料探勘技術來分析病患輸入之資料，而本專題則以此為基礎，建立並比較四種模型：決策樹(decision tree)、隨機森林(random forest)、支持向量機(SVM)、羅吉斯回歸(logistic regression)，並對其進行參數優化，建立更加符合資料集之模型，以得到更加的預測結果。

專業醫師指出，氣喘較有效之觀察天數區間為 3 至 5 天，意即氣喘病患在 3 至 5 天內的氣喘情形會較為相似且可預測，因此若以鄰近天數之氣喘情形、加上天氣及空污資料，並搭配機器學習建立模型用以預測病患之未來氣喘情形，會是十分符合醫學且有效之方法。

五、解決方法

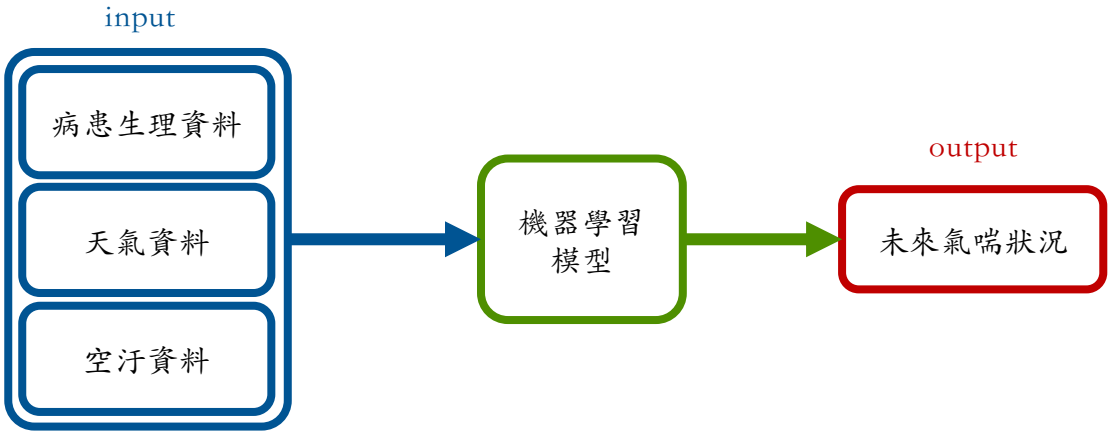
本專題使用病患紀錄的資料作處理，以預測日回推前 1 至 4 天為基礎，篩選並建立有連續天數的資料，將資料導入模型做訓練，再部署至 App 上。當使用者在我們所開發的 App 記

錄完符合連續天數的資料並送出時，系統將自動抓取病患所在地之天氣及空污等相關資料，並結合病患所填的生理資料，向病患展示其未來氣喘嚴重程度並視情況提醒該病患做預防和準備。

本專題目的在於透過讓病患有規律地做日常紀錄，能幫助平時病情的監控，透過模型預測及早得知氣隔日可能之氣喘情形，在面對嚴重的氣喘發作，能先有所準備，降低心理上的不確定感，並告知使用者以做早一步之預防和準備。

六、 系統設計與實作

壹、 模型簡介



貳、 資料介紹

甲、氣喘病患生理資料（來源：台灣南部氣喘照護網，由成大醫院收集）

A. 尖峰呼氣流量值（peak expiratory flow, PEF）：

- 病患以「尖峰呼氣流速計」測量所得，為氣流限制的客觀數據。

B. 病患個人預估值（reference）：

- 病患個人尖峰呼氣流量之正常值。

C. PEFR（peak expiratory flow rate）：

- $\text{PEFR} = \frac{\text{PEF}}{\text{Reference}} \times 100\ (\%$
- 下表指標區間即為模型預測的目標（label）。

綠燈區	PEFR > 80%
黃燈區	80%> PEFR > 60%
紅燈區	PEFR < 60%

D. 各種身體症狀：

- 依據醫學上「one airway, one disease」的概念，同時參考了咳嗽、過敏性鼻炎等其他症狀。

E. 用藥及治療情形：

- 包括氣喘及其他相關疾病。

F. 病患基本資料：

- 年齡、性別等。

乙、天氣資料（來源：中央氣象局）

臺南地區逐日平均氣溫、最高溫、最低溫及平均濕度資料。

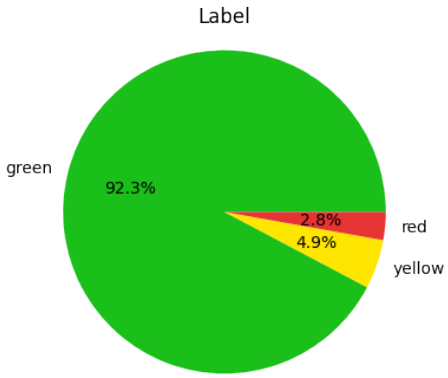
丙、空汙資料（來源：行政院環境保護署環境資源資料庫）

臺南地區逐日各種空污指標物質測量數據。

參、 資料探索與前處理

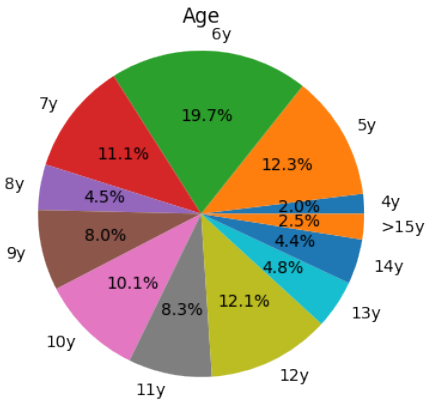
甲、label 統計：

可觀察到資料 label（PEFR）極為不平衡，綠燈區佔了 92%，黃燈與紅燈區僅分別佔了約 5%和 3%。



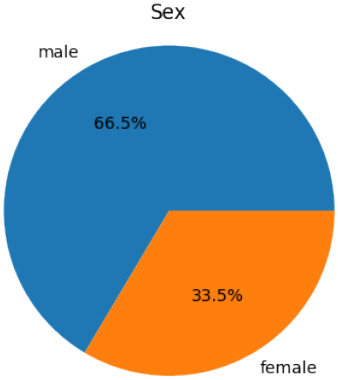
乙、年齡統計：

根據衛福部國健署調查，國人患有氣喘的比率有兩個高峰，主要是在 14 歲前和老年後，觀察資料集也符合此趨勢，14 歲後的資料大幅減少。



丙、性別統計：

根據醫學上的統計，14 歲以前男女患病比例大概是 2:1，此資料集組成以 14 歲前為主，經觀察也符合此比例。

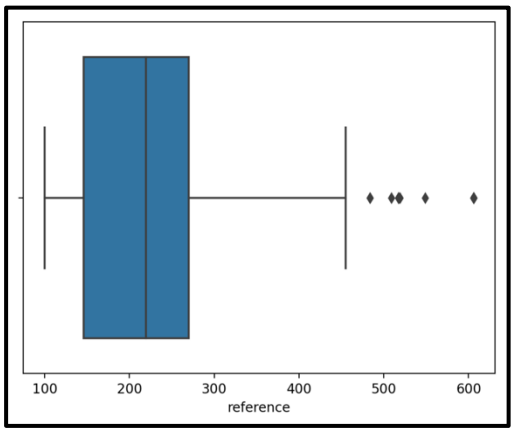


丁、缺失值處理 ：

刪除「氣喘病患生理資料」中有含有缺失值的資料筆數，再將得到的「完整資料筆數」與「天氣和空汙資料」使用日期做 mapping，若對應到的「天氣和空汙資料」含有缺失值，則以缺失值所屬月份的資料平均值做填補。全部資料筆數 4362 筆，刪除 143 筆含缺失值者，得到完整資料 4210 筆。

戊、離群值處理：

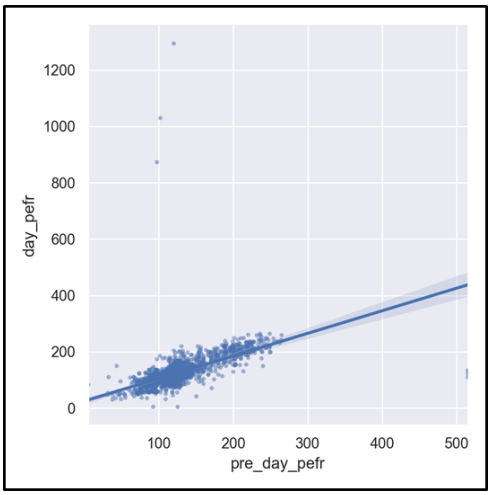
以繪製箱形圖觀察各個特徵，小於第一四分位數（Q1）減去 1.5 倍四分位距（IQR）或大於第三四分位數（Q3）加上 1.5 倍四分位距（IQR）的值視為離群值，再搭配 Domain Knowledge 決定是否刪除該筆資料。完整資料比數 4210 筆，刪除 15 筆含離群值者，得到最後資料 4195 筆。



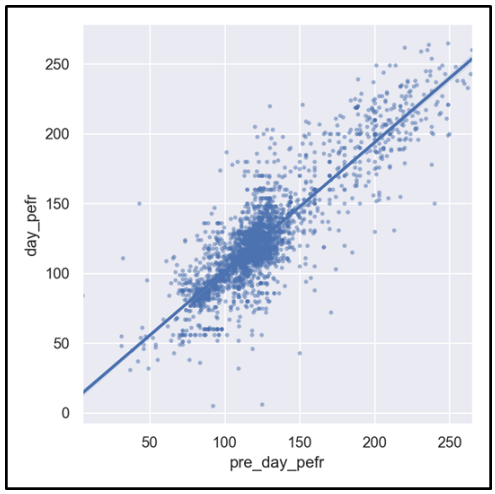
己、觀察離群值影響：

以前一天的 PEFR 值和當天的 PEFR 值做散佈圖，可明顯觀察到離群值的影響。

未處理離群值

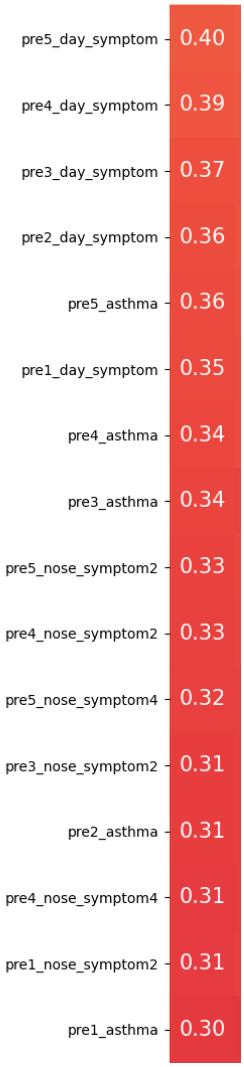


處理離群值後



庚、Pearson correlation：

右圖為當日 PEFR 值與各特徵間的 heat map 部分截圖。可以觀察到，越多天前的相關症狀和服藥情形和當日 PEFR 值相關性越高，推測因為出現相關症狀者會搭配服藥，而服藥天數越多也越能提高 PEFR 值。



肆、 資料特徵工程：

甲、One-Hot 編碼：

處理 Categorical 類別的資料。

乙、Binning：

做數值分段。

丙、新增特徵：

利用原先的特徵，製作新的相關特徵，包括前後兩天的溫差值、濕度差值、各種空污指標差值、早晚 PEFR 變異量等。

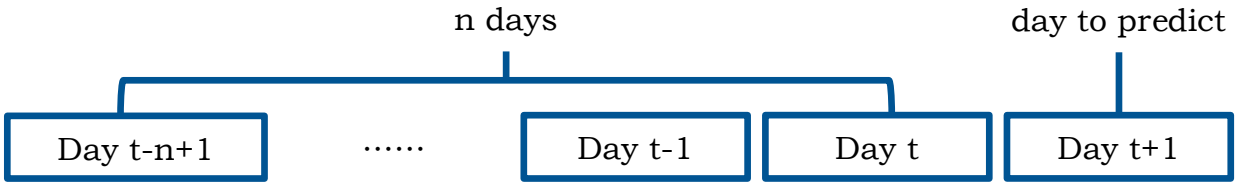
丁、資料降維

在 One-Hot 編碼以及新增特徵過後，資料維度大幅上升，使用主成分分析（PCA），希望能保有原先資料的資訊，並有效的把資料從高維度轉換到低維度。

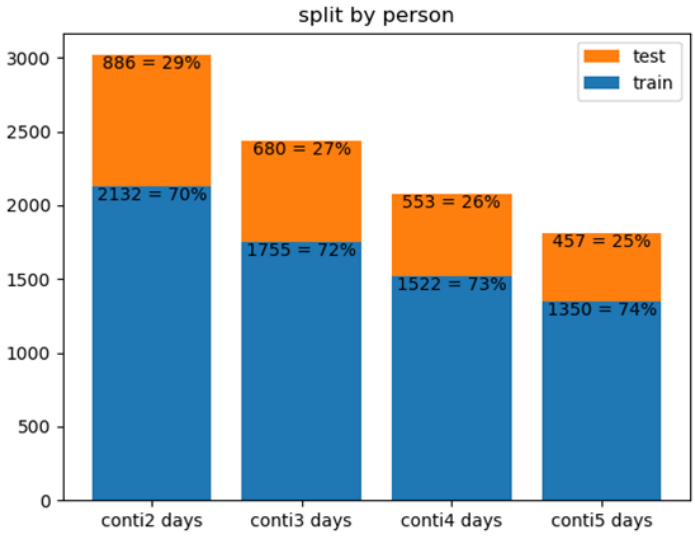
伍、 模型建立

甲、實驗規劃：

為二分類的問題，根據 PEFR 值區間，綠燈區與黃燈區為一類，紅燈區為一類。使用不同的資料集（連續 2~5 天），實驗「前 n 天資料預測下一天」（詳見下圖）當中的不同 n（n=1~4）的設定。不平衡資料處理以未處理不平衡、SMOTE、SMOTE + ENN、調整 weight 參數這 4 種方式進行實驗比較。模型選擇以 Decision Tree、Random Forest、Logistic Regression、SVM 這 4 種模型進行實驗比較。



用「前 n 天資料預測下一天」的氣喘嚴重程度架構圖



不同連續天數資料劃分後的統計

七、成效分析

壹、模型實驗結果：

甲、第一階段：評估指標採用 f1-score

imbalance	n	Decision Tree	Random Forest	Logistic Regression	SVM
None	1	0.517	0.493	0.694	0.493
	2	0.712	0.493	0.829	0.493
	3	0.559	0.494	0.690	0.494
SMOTE	1	0.559	0.594	0.628	0.655
	2	0.606	0.630	0.802	0.637
	3	0.520	0.548	0.658	0.588
SMOTE + ENN	1	0.561	0.671	0.598	0.649
	2	0.580	0.629	0.759	0.634
	3	0.506	0.545	0.654	0.590
Adjust weights	1	0.528	0.493	0.575	0.616
	2	0.598	0.493	0.786	0.613
	3	0.512	0.494	0.654	0.564

未做 imbalance 前，模型表現以 Decision tree 及 Logistic Regression 較佳，做完 imbalance 處理完後，Decision tree 及 Random forest 的表現都不太如預期，推測上述兩項 model，在未做 imbalance 處理前，擅於分出輕症患者，但做完 imbalance 之後則易將輕症患者劃分為重症患者，導致分數下降。

Logistic Regression 在 imbalance 處理過後，雖然表現有稍微下降，但都有維持一定水準，劃分資料的正確性還是足夠，相對其他模型來說，表現都還是滿穩定的。

SVM 在做完 imbalance 處理後的表現有明顯上升， 推測是原本資料中的 minority 在做完 imbalance 處理後有較正確的被劃分，因此分數提高，這也較符合我們的需求 (提高 minority 之預測準確度)。

經過觀察，SMOTE 的表現比 SMOTE+ENN 略加，而兩者都比 weight 調整好上不少，而要決定的天數 n=4 因為在 baseline model 表現即比較有落差，故沒有繼續討論，推測是維度詛咒導致，最後比較 n=1~3，則以 n=2 天的表現最穩定也最好。

乙、第二階段：

根據第一階實驗數據，由於 Logistic Regression 及 SVM 綜合表現也最佳，也能符合本專題的需求，找出較重症之患者，因此決定以這兩種模型進行參數調整。最後的設定：前 n 天資料，n 選擇 2 (用前兩天資料來預測隔天的氣喘情況)，不平衡處理採用 SMOTE，評估指標是 PEFR 紅燈區（label 1）的 recall ，因為本模型希望能從所有 PEFR 紅燈區的情形中，盡可能地找出嚴重情況者並提醒其及早預防，故以此作為最

後優化的評估指標。最後根據實驗結果如下圖，在 f1-score 沒有顯著差異下，選擇 recall 值較高的 SVM 作為最終 app 上部署的模型。

	precision	recall	f1-score	support
0	0.990	0.882	0.933	660
1	0.152	0.700	0.250	20
accuracy			0.876	680
macro avg	0.571	0.791	0.591	680
weighted avg	0.965	0.876	0.913	680

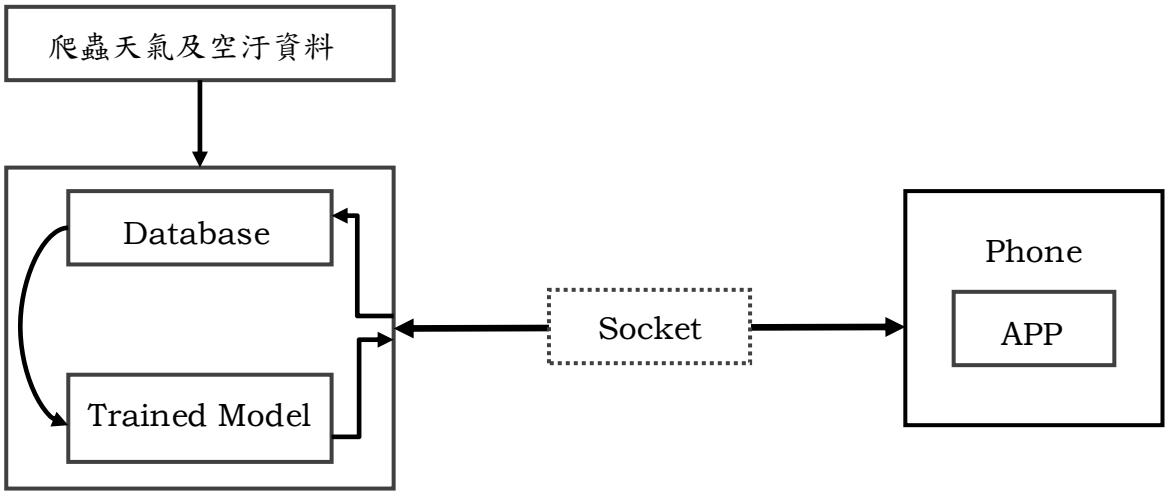
Logistic regression

	precision	recall	f1-score	support
0	0.993	0.856	0.919	660
1	0.144	0.800	0.244	20
accuracy			0.854	680
macro avg	0.569	0.828	0.582	680
weighted avg	0.968	0.854	0.900	680

SVM

貳、App 簡介

甲、架構：



乙、頁面：

A. Home 按鈕：返回 main page

B. 四個主要功能按鈕：

1. 個人資料設定：基本資料填寫。
2. 生理資料輸入：昨日及今日的生理資料填寫。
3. 氣喘情況預測：進行預測並顯示結果。
4. 天氣空汙查看：顯示氣溫、濕度、空污（PSI）指標。

丙、App apk 檔案連結：

八、結論與貢獻

根據我們研讀相關論文[6]的結果，我們認為除了以單一模型分析資料外，可以使用不同模型，並相互比較，在實驗中我們發現，決策樹及隨機森林對於資料中多數類別的預測效果較好，而支持向量機及羅吉斯回歸則在少數類別和綜合的上的表現較佳，由於醫學相關資料的特性，常常會有資料類別不平均的情形，多數為無症狀或輕症患者，少數才是較重症患者，而本專題則是希望能夠找出患者氣喘情況較嚴重的情形，並提醒其及早預防氣喘的發作。

我們的 App 往後將會加入更多功能，例如顯示一整個月的氣喘情況變化，也希望在收集使用者的意見之後能更進一步優化介面，調整模型，另外也希望能夠開發醫師端用的 App，可以直接觀察病患資料情形、做互動或更進一步的預測，希望對氣喘病的預防或協助診斷都能夠更進一步。

九、參考資料

[1] 長庚醫院兒童過敏氣喘中心

https://www1.cgmh.org.tw/chldhos/intr/c4a80air/contents/health01_22.htm

[2] 長庚兒童醫院 歐良修醫師簡報

https://slidesplayer.com/slide/11076571/

[3] 衛生福利部

https://www.hpa.gov.tw/Pages/Detail.aspx?nodeid=1136&pid=3100

https://www.hpa.gov.tw/Pages/Detail.aspx?nodeid=633&pid=1192

https://www1.nhi.gov.tw/mqinfo/Content.aspx?Type=Asthma&List=8

[4] 台中榮總嘉義分院 陳怡成醫師

口述諮詢訪談

[5] Julie L. Harvey 、 Sathish A. P. Kumar

Machine Learning for Predicting Development of Asthma in Children

https://ieeexplore.ieee.org/document/9002692

[6] 陳佳妤

An Integrated Bio-Signal Data Mining Mechanism with Applications on Asthma Monitoring

and Prevention

http://ir.lib.ncku.edu.tw/handle/987654321/21013