2023 年臺灣國際科學展覽會
優勝作品專輯

作品編號　**190021**

參展科別　電腦科學與資訊工程

作品名稱　以機器學習改善罕見疾病之預測

得獎獎項

就讀學校　臺北市立第一女子高級中學

指導教師　林智仁、黃芳蘭

作者姓名　林小凡、陳郁媗

關鍵詞　少標籤、機器學習、多標籤分類

作者簡介

我們是陳郁媗（左）和林小凡（右），目前就讀高中三年級。在這次的專研中我們收穫良多，深入瞭解一個主題難免會遇到挫折，但這個過程中也激盪出更多有趣的想法，除了技術上的知識，在團隊合作與溝通上也有許多收穫。我們很感激一路上得到的幫助，打開了我們的眼界。我們很高興可以有這次機會，豐富了高中三年的生活，也多了很多難忘的回憶。

摘要

我們想利用機器學習進行疾病診斷，但現有方法對罕見疾病的預測精確率低，且若過於專注在罕見疾病預測的提升，容易導致整體精確率降低。

為了兼顧整體與罕見疾病的精確率，我們將預測分為兩個階段。在第一階段運用現有的多標籤分類方法訓練，第二階段使用二元成本導向判斷病人是否有罕見疾病，再利用第二階段得到病人有無患有罕見疾病的結果，決定是否在預測此病人可能患有的疾病時，保留一個位子給罕見疾病。實驗結果呈現在兩階段皆用神經網路（Neural Network, NN），能正確預測罕見疾病的比率為現有方法的八倍，而整體精確率只下降 0.02，並實作出疾病預測系統。

Abstract

Existing multi-label classification for disease prediction encounters problems in some challenging scenarios. For instance, rare diseases are often hard to predict. These rare diseases correspond to few-shot labels in multi-label classification, and the rare occurance means that focusing too much on their performance may lead to a decrease in the precision overall. Therefore, our project aims to find a solution that can raise the precision of these few-shot labels while maintaining the overall precision.

We divide the prediction into two phases. In the first phase, we apply standard multi-label training. In the second phase, we conduct post-processing to use binary cost-sensitive learning for deciding if a patient is diagnosed with rare disease. This result determines if we leave a spot for rare disease when predicting the diseases this patient is diagnosed with.

壹、研究動機

曾經看過新聞有病人因為疾病被誤判，因此腎臟被切除而影響終生。這樣的事件導致人們在就醫想先給有經驗的醫生診斷疾病。但畢竟人力有限，不可能所有病人都如願以償能找到有經驗的醫生。況且，許多偏遠地區的居民，要到大醫院須費盡千辛萬苦，容易導致疾病太晚被診斷出來，而失去治療的最佳時機。

若能將醫師的診斷病人的大量資料透過機器學習進行疾病分析，是否能使醫療體系更加完善？

目前根據病人資料進行疾病診斷及判斷對應的國際疾病分類（International Classification of Diseases, ICD）這些項程序是由醫師進行。若發展出疾病預測模型，就能節省人力。但這些模型在預測疾病時，常常忽略罕見疾病，導致罕見疾病無法被預測。因此若能改善模型對於罕見疾病的預測，能幫助醫生做出更正確的判斷。

貳、研究目的及研究問題

一、比較現有的罕見疾病預測模型。

二、嘗試解決罕見預測結果不理想的問題。

三、提出能兼顧罕見疾病預測與整體表現的演算法並實作系統。

參、研究設計

一、疾病分類

醫生看診時，會將病人的症狀與資料紀錄，而這些資料就成為判定病人患有疾病的依據。透過大量的數據，可以設計出能預測疾病的模型。病人的資料為文字敘述，而預測的結果為對應到國際疾病分類的疾病的數字，數字有時不只一個，同一位病人可能同時有多種疾病。因此這樣的分類屬於文本多標籤分類。在製作模型時，先對文字資料進行處理，而後進行預測，每個疾病會有相對應的機率，表示此病人患有這個疾病的可能性。由人為定一個數字 K，最後顯示出前K個可能的疾病。

若以疾病作為標籤，罕見疾病即為少標籤。但預測模型在少標籤的預測，常會表現不理想。又少標籤出現機率少，即使忽略它，整體準確率仍然能達一定標準，難以發現少標籤預測正確率低的問題，而模型無法預測少標籤的問題，也導致無法取代人力來進行疾病分類。

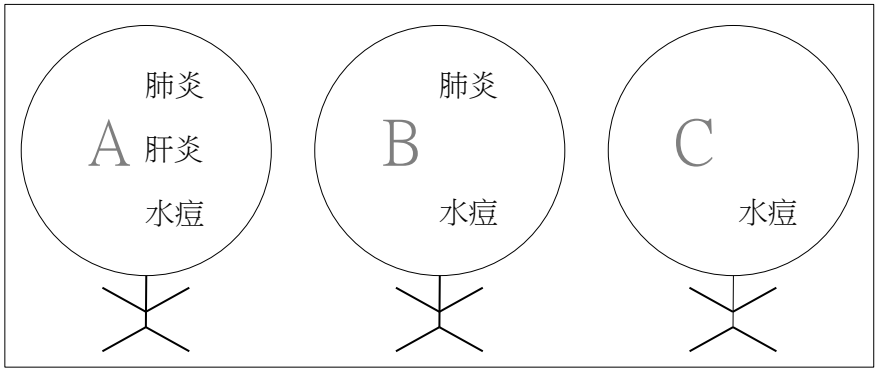
少標籤在最後的預測結果幾乎不會出現在前面 K 個。導致在進行診斷時容易忽略罕見疾病，因此我們希望能夠改進罕見疾病的預測。但若過度追求罕見疾病預測的表現，可能導致其他較常出現疾病的預測結果不如以往，導致整體預測的準確率下降。所以我們希望能在罕見疾病預測與整體表現取得平衡。找到兼顧整體準確率與少標籤預測的模型，為疾病分類領域貢獻心力。

二、文獻回顧

（一） 多標籤分類（Multi-label Classification）

病人的資料稱作實例（instance），疾病稱作標籤（label），一個病人可能同時有多個疾病，相當於一個實例會同時分類到多個不同的標籤，此分類問題稱作多標籤分類問題。以下圖為例，假設有三位病人（A、B、C）可能有這三種疾病的其中一個或是多個。A 病人同時患有三種疾病，而 B 病人患有肺炎及水痘兩種疾病， C 患者則是患有水痘。以 1 代表病人患有此疾病，0 則代表此病人不具有

此疾病。則病人 A 的表示法為 [1, 1, 1]，病人B的表示法為 [1, 0, 1]，病人C的表示法為 [0, 0, 1]。

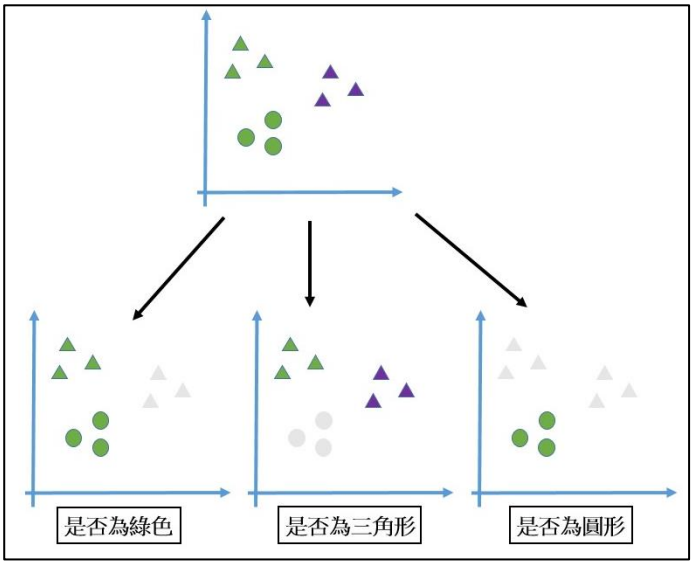


圖一、多標籤分類範例圖

由此可知多標籤分類並非將每個被分類的實例分配到一個標籤，而是找出所有它屬於的標籤。

由於大多數的演算法只能處理單標籤分類，所以處理多標籤分類問題時，多半會將其分解成好多個單標籤分類問題，如此一來，現有的處理單標籤分類問題的演算法才能被使用。

One vs. Rest 是一種處理多標籤問題的對策，在分類的時候，以某個類別的標籤分為一類，其餘的分為一類。若要分的類別有n個，則重複這個動作n次，即會 n 次的把每個類別與其他類別分開，以得到每個實例屬於哪些類別。下圖為 One vs. Rest的例子。



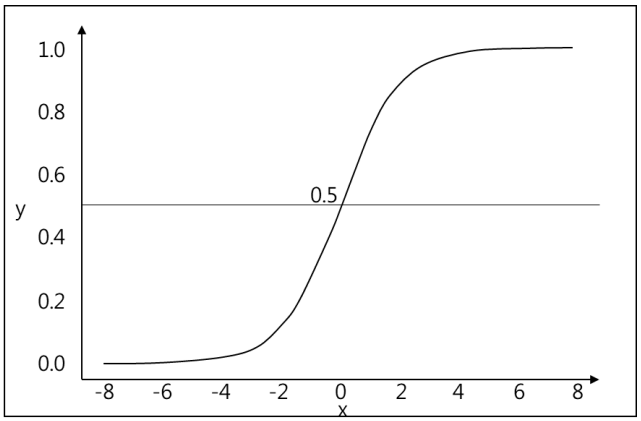
圖二、One vs. Rest範例圖

上面這個例子中，我們總共有三個類別需要分。先判斷樣本是否為綠色，再判斷是否為三角形，最後判斷是不是圓形，每個條件都能區分出一個類別。

而One vs. Rest又可以透過邏輯迴歸與神經網路的方式來執行。

1. 邏輯迴歸（Logistic Regression, LR）

使用邏輯回歸進行 One vs. Rest 的分類時，對於每個標籤進行判斷是或不是。而判斷是或不是則是由方程式所決定，方程式得到的值為0到1之間，得到的值即為機率，這時會設定一個定限（threshold），機率高於或低於定限即為判斷是否屬於此標籤的依據。



圖三、邏輯分布函數圖

例如判斷圖案是否為三角形，若定限為 0.5，則機率超過 0.5，判斷為三角形，反之不為三角形。舉例來說，若得到的機率為 0.7，則判斷圖案為三角形。

2. 神經網路（Neural Network, NN）

假設多標籤分類問題中有n個標籤，設計一個神經網路有n個判定函數。對於每個實例，進行n個判斷，就能夠得到此實例屬於那些標籤。

（二）少標籤（Few Shots）

對於一個資料集中的標籤，有些出現次數頻繁，有些則出現次數少。根據資料的特性，我們可以定義在資料集中出現少於一個固定次數的標籤為少標籤。若不同的標籤代表不同的疾病，則少標籤即為罕見疾病。

疾病預測就是將一個病人可能得到的疾病作排序，排序在前面的即是病人最可能患有的疾病。常見的疾病較容易被排在預測出的前幾名，而罕見疾病被排序在前的機率非常小。假設今天一位病人患有急性中耳炎、過敏性鼻炎、氣喘與乙醯穀胺酸合成酶缺乏症，其中最後一項為罕見疾病，則很難預測出此病人患有乙醯穀胺酸合成酶缺乏症。

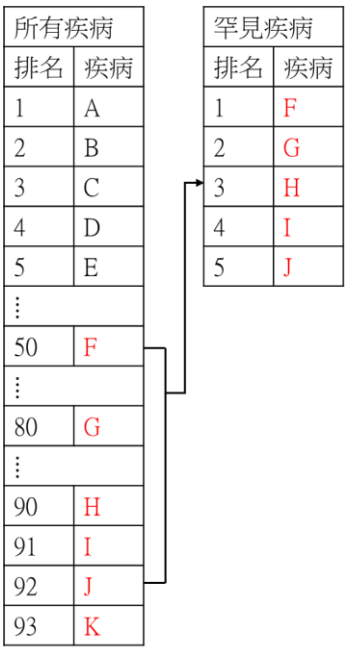
少標籤出現次數少，因此可能不會對多標籤模型的整體準確率造成很大的影響，但罕見疾病不能被忽視，因此不能忽略少標籤在模型中的分類。

（三）現有方法疑點

在相關領域中研究少標籤的文獻裡，Rios and Kavuluru (2018) 較具代表性，被很多人引用，且提出改善少標籤預測的方法。

在疾病預測中，會根據病人的症狀給予醫生一列病人可能患有的疾病的排名。罕見疾病為少標籤，較難被預測，對於整體預測的表現影響也較小，因此評估少標籤預測的表現也較困難。此研究在進行預測後，對於罕見疾病表現的評估，是將少標籤，分開獨立排名。依據病人最可能患有的罕見疾病排名。因此一位病人會有兩列可能患有的疾病，其中一列包含所有的疾病，另外一列只有罕見疾病。而罕見疾病預測的表現的評估方式就只看其在只有罕見疾病的排名中的表現。

例如： 設定對於一位病人顯示前5個可能的疾病，今天有一位病人患有疾病 A, B, C, D, F，其中F為罕見疾病。假設總共有 100 種疾病。因為F為罕見疾病，因此很難被預測在所有疾病中的前五個。前五名可能為 A, B, C, D, E，皆為常見疾病。如右圖所示，以紅字表示罕見疾病，假設罕見疾病在所有疾病中的排名在 50, 80, 90, 91, 92, 93位，若將罕見疾病獨立排名，則在所有疾病中排名 50 的疾病在罕見疾病排名中第1，排名 80 的為 2，排名 90 的為 3，依此類推 。



圖四、疾病預測評估方法示意圖 [1]

此研究對於罕見疾病的表現評估如下：

1. 經多標籤分類模型的預測後，對每位病人會產生一個所有疾病的排名與只含有罕見疾病的排名。
2. 對於罕見疾病表現的評價，會單獨從罕見疾病排名內取出前K個做比較。

在這樣的設定下，罕見疾病的表現大幅提升，下表為實驗的部分結果。

表一、所有標籤與少標籤的Recall [1]

所有標籤	少標籤
R@5	R@5
0.135	0.130

其中，所有標籤代表的是所有的疾病，包含常見的疾病與罕見疾病。少標籤則代表罕見疾病。 R@5 代表顯示前 5 個疾病時的 Recall，關於 Recall將在研究過程或方法中詳細介紹。由此表可以得知，預測前 5 個疾病時，所有疾病的召回率為 0.135，罕見疾病的召回率為 0.130。

可以從上述的數值中發現，所有疾病的表現與罕見疾病的召回率相近。罕見疾病在預測前 5 個疾病時因為評斷罕見疾病預測的表現時，只看罕見疾病的排名，所以能夠達到與整體相近的召回率。

上述少標籤的結果是只看罕見疾病排名時的結果，此結果在現實中無法表現少標籤預測的全貌，病人患有的罕見疾病即使在所有罕見疾病中排名最前面，若在所有疾病中排名很後面，最後也無法被顯示出來。因此上面少標籤的結果過於樂觀，在現實中少標籤的預測並無法達到這樣的結果。

現有方法問題：
由於罕見疾病被分開排名，所以即使罕見疾病的預測正確率提高，也不代表其在所有疾病中的預測有所改善。實際上在所有疾病當中罕見疾病仍然處於被忽略的狀態。

三、研究設備及器材

（一）硬體工具 ：

1. CPU ： Intel（R） Core（TM）i5-10210U CPU@1.60GHz 2.11GHz
2. 記憶體 ： 12.0 GB
3. 作業系統 ： Windows 10

（二）軟體 ：

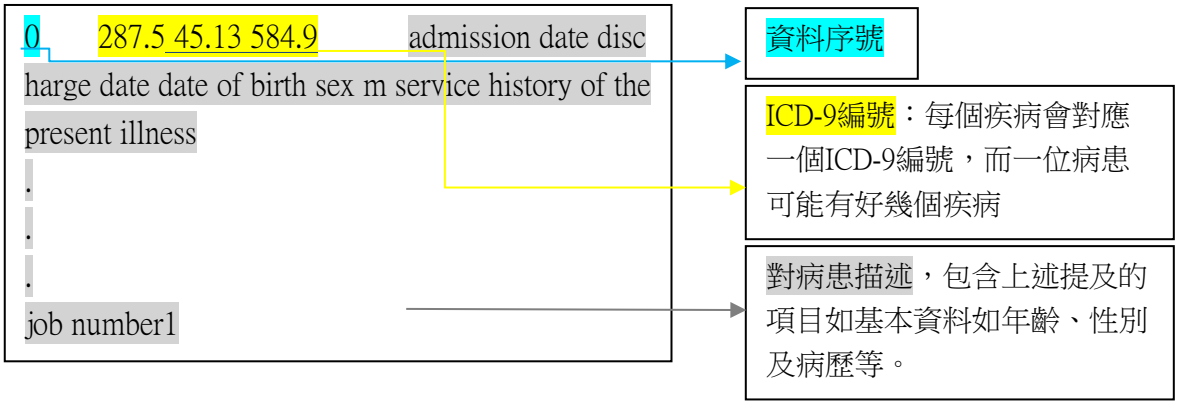
1. Matlab R2021a
2. Python（3.7）
3. Pytorch（1.8）
4. LibMultiLabel

LibMultiLabel 是一個處理多標籤文件分類 （Multi-label Text Classification）的機器學習開源軟體。該工具可在https://github.com/ASUS-AICS/LibMultiLabel下載

（三）訓練資料 ：

我們使用的訓練資料為MIMIC-III Clinical Database。

MIMIC-III 是一個大型免費醫療數據庫，紀錄四萬多名在2001至2012年間住過貝斯以色列女執事醫療中心（Beth Israel Deaconess Medical Center, BIDMC）重症監護病房的患者健康相關資料，包含基本資料如年齡、性別及病歷，與每小時一次的臨床生命徵象測量結果、實驗室測試結果、藥物、護理人員筆記、影像報告和死亡率。以其中一筆資料為例： 0 為資料序號，287.5 45.13 584.9為 ICD 第九版（ICD-9）的編號，admission date discharge date date of birth sex m service history of the present illness…job number1 為對病患的描述。

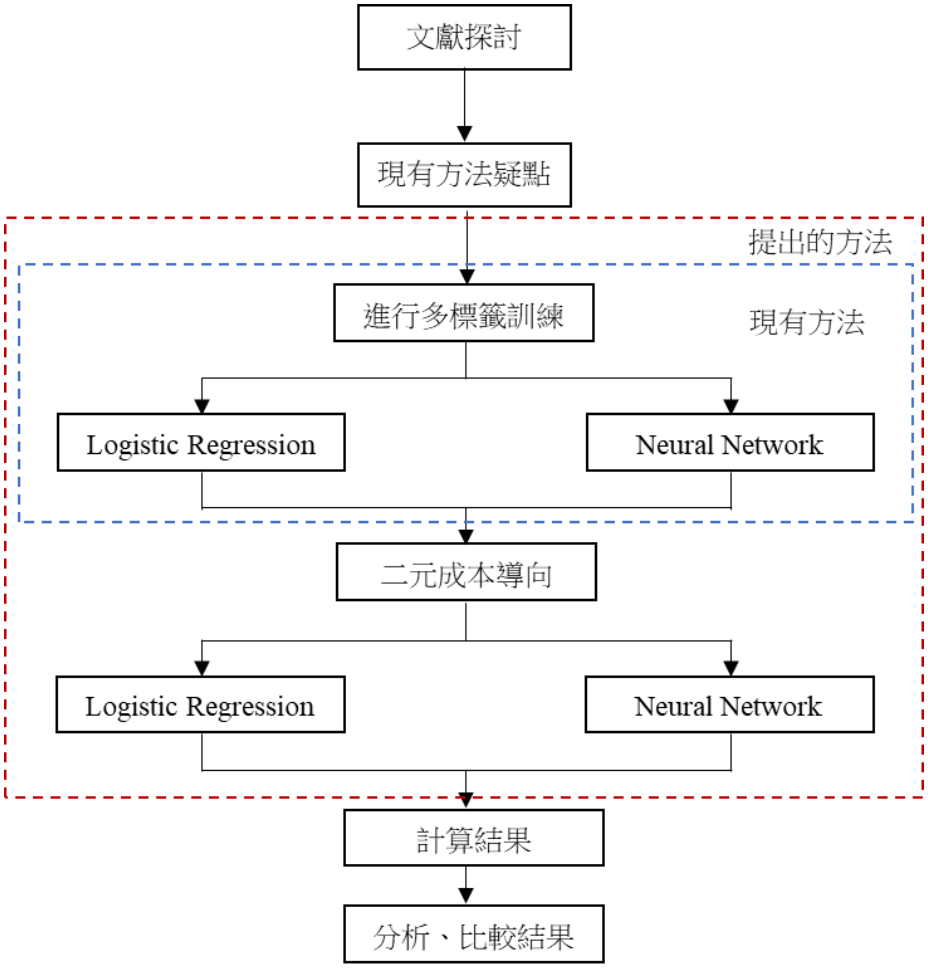


圖五、ICD-9資料示意圖

表二、ICD-9編碼列表（擷取）

001	霍亂（虎烈拉）
001.0	霍亂弧菌所致的霍亂
001.1	異性霍亂弧菌所致霍亂
001.9	霍亂
002	傷寒及副傷寒
002.0	傷寒
002.1	A型副傷寒
002.2	B型副傷寒
002.3	C型副傷寒
002.9	副傷寒

四、研究流程



（一）訓練資料

MIMIC-III，如上面所述。

（二）現有方法疑點

1. 邏輯迴歸（Logistic Regression）

疾病分類問題中，由於讀入的資料為文字，因此須將文字經過 TF-IDF的處理後轉為數字，一個詞彙 s 的詞頻與逆向文件頻率乘積為此詞彙的 TF-IDF。以下為算式及說明。

$$TF - IDF(s) = TF(s) \times IDF(s)$$

（1）詞頻（Term Frequency）

某一個給定的詞語在一文件中出現的頻率。若第 s 個詞在總字數為 n 的文件中出現 t 次，則詞頻為 $\frac{t}{n}$。

$$TF(s) = \frac{t}{n}$$

（2）逆向文件頻率（Inverse Document Frequency）

以處理常用字問題。若詞彙 s 共在 d 篇文章中出現過，而總共有 D 篇文章，則詞彙 s 的逆向文件頻率為 $\log\frac{D}{d}$。

$$IDF(s) = \log\frac{D}{d}$$

若一個詞彙s在文章d中出現頻率高，且很少出現在其他篇文章，則 TF-IDF 整體來說較高。藉由 TF-IDF 可以評估一個詞彙對於一個文件的重要程度。

得到病人資料中詞彙的 TF-IDF 後，可以得到病人的向量。向量長度為字典詞彙的數量，對於每個詞彙即為對應的 TF-IDF，未出現的詞彙即為0。

得到病人的向量後，由疾病的模型判斷病人患有此疾病的機率。每種疾病都有一個機率模型。

機率模型（Probability Model）

$$p(y|x) = \frac{1}{1 + e^{-yw^{\mathrm{T}}x}}$$

- y： 1或-1，代表有此疾病與不具有此疾病
  - $p(1|x) + p(-1|x) = 1$
- $w^{\mathrm{T}}x$ ： w 向量與 x 向量之內積
  - w： 參數 （Weight Parameter）
  - x： 代表病人之向量

對於一個疾病的機率模型，需要經過訓練，使概似函數（Likelihood Function）最大值，得到此疾病的w向量。

概似函數（Likelihood Function）

$$\prod_{i=1}^{l} p(y_i|x_i)$$

- 對於一個疾病，將每為病人經機率模型得到機率相乘
  - l：病人總數
- 訓練時希望能使概似函數最大
  - 例：
    兩位病人分別具有及不具有此疾病,兩人經過機率模型得到的機率如下
    - $p_1\left(1\middle|x_1\right)$
    - $p_2\left(-1\middle|x_2\right)$
  - 此疾病的 w 向量期望能使兩者機率最大,因此使兩者乘積最大化。

由於病人總數繁多，機率相乘後乘積太靠近 0，因此改以負對數似然（Negative log-likelihood）訓練，使其最小。

負對數似然（Negative log-likelihood）

$$-log\prod_{i=1}^{l} p(y_i|x_i) = -\sum_{i=1}^{l} \log\left(1 + e^{-y_i w^{\mathrm{T}} x_i}\right)$$

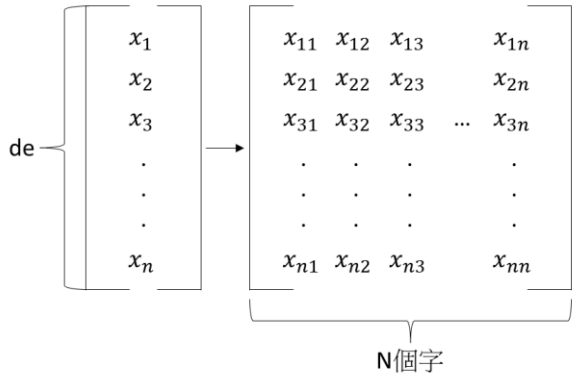
- l 是訓練資料的數目

（二）CAML 神經網路（Convolutional Attention for Multi-Label Classification）

神經網路分為以下幾個步驟。

1. 卷積（Convolution）

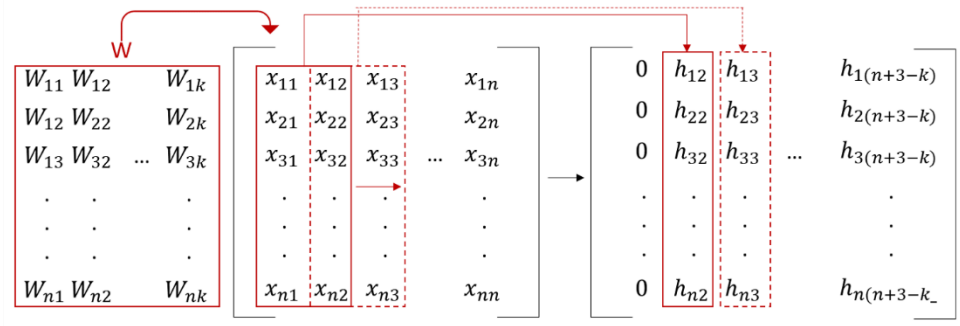
一位病人資料中，每個字都有詞嵌入（word embedding）成為長度為 d_e 的向量。然後將每個字的向量排列，得到 $d_e \times N$ 的矩陣 X，其中 N為資料長度。



圖六、文字以向量表示

接著使用大小為 $d_e \times k$ 的濾波器（filter）W，進行卷積後得到向量 h。為使 h 之長度為 N，先將 X 進行墊零（zero padding）。

總共有 d_c 個濾波器，則經過卷積後的向量排列後將得到 $d_c \times N$ 的新矩陣 H。



圖七、卷積與墊零

卷積運算

$$h_n = g\left(W_c * x_{n:n+k-1} + b_c\right)$$

- *：將 W_c 矩陣與 $x_n \sim x_{n+k-1}$ 陣列所對應的值相乘再加總
- g 做非線性的轉換，b_c 為偏值（bias）。

2. 注意力（Attention）

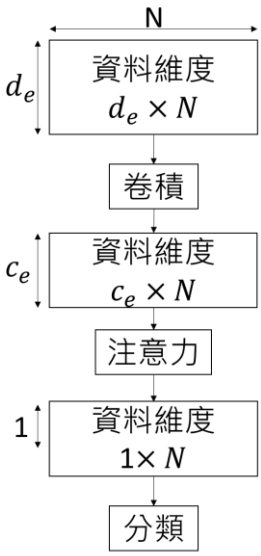
每種疾病都有一個代表參數向量 u_l，長度為 d_c，和所有 h_n 進行矩陣乘法得到向量再進行 Softmax 後為 α_l。

Softmax

使一組向量中的數字皆位於（0,1）之間，且加總後為1。因此將每個數字經指數轉換，再除上所有經指數轉換的數字的加總。

假設T為加總

$$\begin{matrix} a & e^a & e^a/T \\ b \rightarrow & e^b & \rightarrow e^b/T \\ c & e^c & e^c/T \end{matrix}$$



圖五、神經網路步驟

接著將 α_l 向量對每一列使用則代表此疾病的向量為 v_l

$$v_l = \sum_{n=1}^{N} {\alpha_l}^T h_n$$

- N為文獻長度

3. 分類（Classification）

對於一位病人的向量，使用Linear Layer與Sigmoid，計算患有疾病l的機率。

Sigmoid

$$\sigma(x) = \frac{1}{1+e^{-x}}$$

屬於一個激發函數（activation function），輸出介於0到1之間的數字。

患有疾病的機率

$$\hat{y} = \sigma({\beta_l}^T v_l + b_l)$$

- β_l 為權重向量
- b_l 為偏值向量

4. 訓練（Training）

調整 β_l 使二值交叉熵（binary cross-entropy loss）損失最小

$$L_{BCE}(X,y) = -\sum_{i=1}^{L} y_i \log(\hat{y}_i) + (1-y_i)\log(1-\hat{y}_i)$$

- BCE 代表 Binary Cross-Entropy Loss

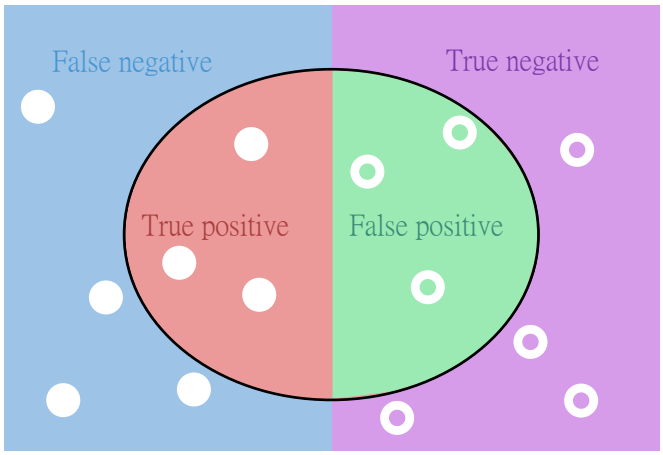
（二）資訊檢索評價

預測出疾病後，需要評估模型的預測結果與實際值的誤差，以修正模型達到更高的準確率，我們利用以下方法或步驟進行評價。

1. 混淆矩陣（Confusion Matrix）

表三、混淆矩陣

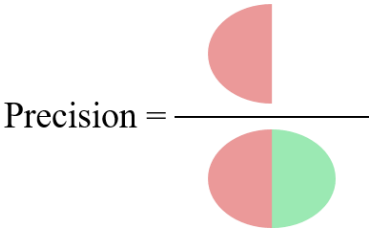
混淆矩陣		真實類別	
		有病	沒病
預測類別	有病	True positive（TP）	False positive（FP）
	沒病	False negative（FN）	True negative（TN）



圖八、混淆矩陣示意圖（每一個點：實圈表示Positive，虛圈表示Negative。黑色大圈內之點為被預測為Positive。）

2. 精確率（Precision）

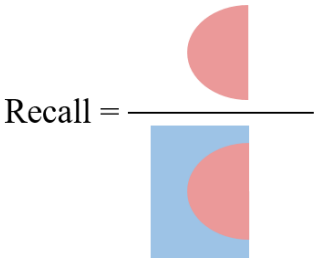
在所有預測類別為真的樣本中，有多少真實類別為真。若精確率高，表示該模型將真實類別為假的樣本預測成真的機率小。



$$Precision = \frac{TP}{(TP + FP)}$$

3. 召回率（Recall）

在所有真實類別為真的樣本中，有多少預測類別為真。若召回率高，表示該模型將真實類別為真的樣本預測成假的機率小。



$$Recall = \frac{TP}{TP + FN}$$

4. P@K及 R@K

P@K是Precision@K的簡稱，R@K是Recall@K的簡稱。我們把預測的$\hat{y}$做排序，選出最前面K個為此實例（病人）對應的標籤（疾病）。此K個標籤即為我們對Positive 的預測（因此k=TP+FP）。算出的精確率及召回率即為P@K與 R@K。因此

$$\text{P@}k(\hat{y}, y) = \frac{1}{k}\sum_{i \in rank_k(\hat{y})} y_i$$

5. Propensity Weighted Precision（PSP）

這是P@k的延伸，可用來評估少標籤預測的結果，簡稱PSP。依據標籤出現的次數給予適當的權重，出現次數少的標籤給予較高的權重，反之則給予較低的權重。

$$\text{PSP@}k(\hat{y}, y) = \frac{1}{k}\sum_{i \in rank_k(\hat{y})} \frac{y_i}{p_i}$$

關於 p_i 的算法，我們參考Bhatia et al. (2016)。

6. F1 Score

精確率與召回率的調和平均數，用它來計算模型的精確度，公式如下。

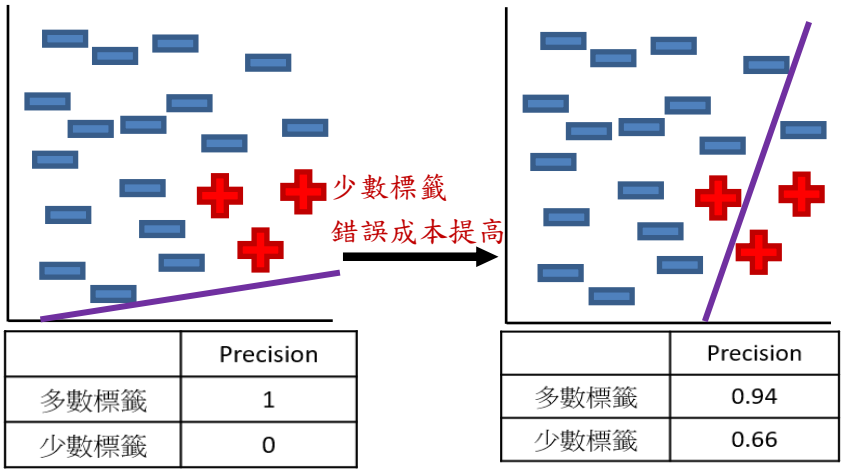
$$\text{F1} = \frac{2 \times Precision \times Recall}{Precision + Recall}$$

四、研究改良方法

在現有方法疑點中可以看出，就算一個方法能在預測罕見疾病時，成功將患有的罕見疾病排在其他罕見疾病前面，排在最前面的罕見疾病仍然無法在所有疾病中得到夠高的排名，所以無法顯示給醫生。因此這樣的方法無法確實反映罕見疾病真正的表現好不好，為了改善此狀況，我們從成本導向（Cost-Sensitive Learning）的角度探討少標籤預測。

(一) 成本導向（Cost-Sensitive Learning）

在分類的過程中會有些錯誤的分類，而成本導向即是對於不同的錯誤分類有不同的對待方式。若一個問題中大部分的實例為負值而只有少數正值，則為正值的實例佔少數，因此對於整體的影響不大，若將全部的實例都預測為負值，也能得到不錯的結果。一個能避免所有實例都被預測為負值的方法就是將正值預測錯誤的成本提高。在機器學習中，成本導向就是一種能處理如上述不平衡分布問題的技術。



	Precision
多數標籤	1
少數標籤	0

	Precision
多數標籤	0.94
少數標籤	0.66

圖九、成本導向示意圖，其中 Precision 的算法為被預測正確的個數除以真實類別個數

(二) 少標籤與成本導向應用

前面提到多標籤分類顯示出最前面的疾病多半是常見的疾病，這樣的情況類似於上述負例占多數的情況，因此能將成本導向的概念應用於此。將少標籤指定較大的成本，預測錯誤時需付出較大的代價。換句話說，提高少標籤的權重，能使少標籤對於整體表現有較大的影響。

因為疾病分類為多標籤分類問題，一個實例並不是只用考慮對於一個標籤的分類結果，而是很多的標籤，因此一般會將一個實例對於所有的標籤的損失

加總。損失加總即為將實例對於所有標籤的損失加起來。從疾病分類的角度來看，就是將一位病人對於所有疾病的分類結果的損失加總。

$$-\sum_{i=1}^{L} y_i \, log\hat{y} + (1 - y_i) \log(1 - \hat{y}),$$

其中 L 是標籤的總數，且 y_i 等於 0 或 1是代表真實的標籤：如果實例與標籤 i 相關聯，則 $y_i = 1$，否則 $y_i = 0$。此外，由模型決定的 P_i 是介於 0 與1中間的數，它是關於實例是否與標籤 i 相關聯的概率估計。

而將其加上不同的成本則為以下的公式。

$$-\sum_{i=1}^{L} C_i(y_i log\hat{y} + (1 - y_i) \log(1 - \hat{y})),$$

其中 C_i 是標籤 i 的權重。以疾病分類的角度來看就是疾病 i 的權重。

每調高不同疾病的權重，會得到不一樣的結果。若不調高罕見疾病的成本，罕見疾病對於整體的影響過小，當整體的精確率提升時，罕見疾病的排名仍然無法被顯示給醫生。若將罕見疾病的成本提高，罕見疾病有比較大的影響力，因此使用成本導向進行預測，可以改進罕見疾病的預測。但調高罕見疾病權重的同時，其他疾病的權重也會降低，因此若過度提升罕見疾病的表現，可能會導致其他疾病的表現下降，而其他疾病仍然占多數，所以整體的表現會下降許多。

表四、使用成本導向的優缺點

使用成本導向：	
優點	罕見疾病表現提升
缺點	整體預測表現下降

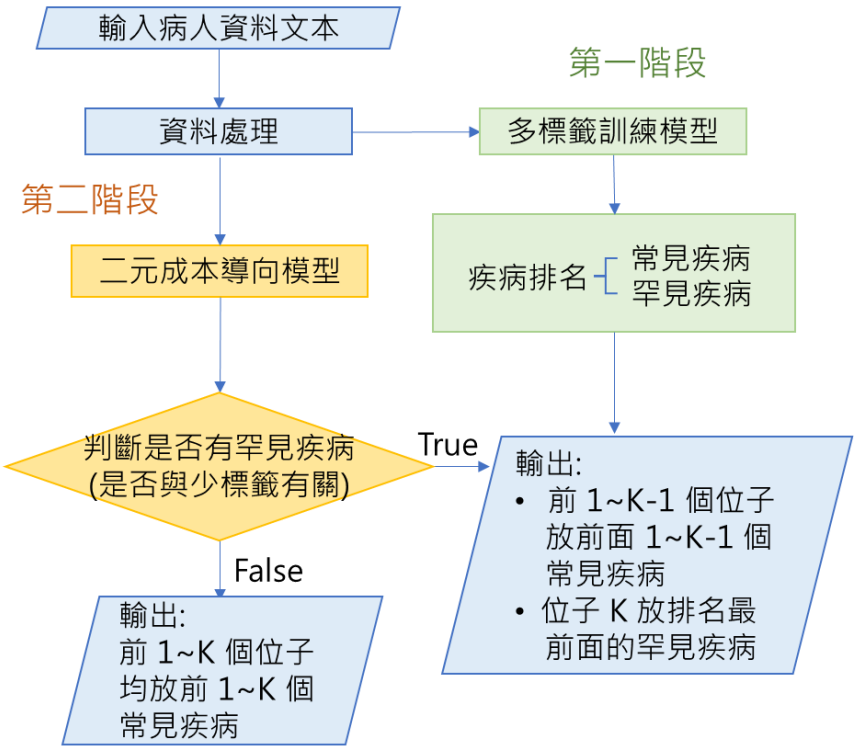
若將整個預測過程都使用成本導向，會造成整體結果下滑，若不使用成本導向，罕見疾病的排名不夠前面。需要一個方法，使少標籤的改善與整體表現的維持能達到平衡。但每個疾病都有不同的 C_i，因此很難對每個疾病都進行調整權重。

為了達到罕見疾病預測的改善，且同時維持整體預測的水準，我們將預測階段分成兩個部分。

在第一階段運用現有的多標籤分類方法訓練，不論這個方法有沒有對於罕見疾病有做特別的處理。而第二階段簡化上述的成本導向使其成為二元成本導

向，將有罕見疾病的病人與不具有罕見疾病的病人賦予不同的權重，判斷一位病人是否具有罕見疾病。此部分將在改良方法實作中有更詳細的解釋。

接著在預測階段，若一位病人在第二階段的判斷為具有罕見疾病，則顯示給使用者的可能患有疾病若有K個，會將最後一個位子留給最可能患有的罕見疾病。若在第二階段的判斷為不具有罕見疾病 ，則顯示的K個疾病不會多做處理。



圖十、改良方法流程圖

五、改良方法實作

現有方法未對罕見疾病多做處理，Rios and Kavuluru (2018) 在改良預測模型時多注重在整體表現的提升，因此罕見疾病預測仍然無法有明顯的進步。我們提出將預測分成兩個部分，以改善罕見疾病預測。

(一) 資料處理

使用Mullenback et al. (2008) 文獻所做的處理

1. MIMIC-III資料庫有26個表，包含醫生類型、處方、診所等。我們使用NOTE EVENTS表中的Discharge Summaries，將關於一位病人的敘述濃縮到一份資料中。
2. 記號化（tokenization）

(1) 以空白分開。
(2) 每份資料最多2500個token。
(3) 不包含英文字母則移除。
(4) 將所有字母轉為小寫。
(5) 出現在少於三分資料的token改為UNK token。

3. 將字詞轉為向量形式

(二) 第一階段： 以現有方法進行多標籤訓練

現有的多標籤分類方法中，具代表性的為線性分類器與神經網路。

1. 邏輯迴歸（Logistic Regression, LR）

邏輯迴歸是一種二元分類的方法，但是我們延伸它進行多標籤分類的問題。這種方法稱為One vs. Rest。就是每個疾病都有一個二元分類問題，判斷所有的病人有沒有患有此疾病。

2. 神經網路（Neural Network, NN）

此處我們使用Libmultilabel的多標籤分類工具神經網路的模板。並且使用Mullenback et al. (2008) 中提出的CAML神經網路。

Libmultilabel的資料為以下所示。

```
2286<TAB>E11 ECAT M11 M12 MCAT<TAB>recov recov recov recov excit ...
2287<TAB>C24 CCAT<TAB>uruguay uruguay compan compan compan ...
```

- 一行代表一位病人
- 由左到右分別為編號、患有疾病、病人相關敘述，由<tab>鍵分開
- 疾病由空白鍵分開

我們參考Liu et al. (2021) 這篇文獻中對不同的參數設定下，Neural Network的結果，並根據不同參數在P@5時的表現選擇最後使用的參數。參數的定義在前面文獻探討中解釋。

表五、參數與定義

參數	定義
#filters	卷積層中的濾波器數量
filter size	濾波器為d×w，其中d為代表文字的向量大小，w為filter size
dropout prob.	在進行卷積前將一些文字的向量丟棄，丟棄的機率。
learning rate	模型的學習率

表六、Liu et al. (2021) 文獻中結果

參數選擇				結果
#filters	filter size	dropout prob.	learning rate	P@5
550	6	0.6	0.0001	0.6426
550	6	0.4	0.0001	0.6356
550	6	0.8	0.0001	0.6353
550	8	0.6	0.0001	0.6348
450	6	0.6	0.0001	0.6347

表七、最終使用的參數

參數	最終使用
#filters	550
filter size	6
dropout prob.	0.6
learning rate	0.003

由表六與表七可以看出P@5可以表現最佳時為0.6426，我們選擇達到此結果的參數作為最終使用之參數。

(二)第二階段： 以二元成本導向判斷病人是否具有罕見疾病

1. 二元成本導向（Binary Cost-Sensitive Learning）

前面提過，在機器學習中，成本導向就是一種將少數的預測錯誤成本提高，以處理不平衡分布問題的技術。

但疾病分類中疾病數量龐大，有上千種疾病，若對於每個不同的疾病都給予不同的成本，會相當複雜。由於我們的目標在於改善罕見疾病（少標籤）的預測，我們著重在常見的疾病與罕見疾病間的平衡。因此我們在第二階段使用二元成本導向，將有罕見疾病的病人與沒有罕見疾病的病人分別給予兩種不同的成本。最後預測疾病顯示給使用者時將使用到第二階段得到病人是否具有罕見疾病的結果。

其中具有罕見疾病與不具有罕見疾病的權重分配如下

(1) 具有罕見疾病： 權重以C+ 表示。
(2) 不具有罕見疾病： 維持原狀，即權重為1。

任何二元分類的機器學習方法在這裡都可以被使用。那我們考慮與在第一階段已被使用的兩種方法。這兩個階段使用的機器學習方法可以每有任何關連，所以舉例來說，我們可以在第一階段使用邏輯迴歸，而第二階段使用神經網路。

(1) 邏輯迴歸（Logistic Regression, LR）

在第二階段中的邏輯回歸我們仍然使用具有定限的邏輯回歸，而對於成本的決定，我們選擇5作為C+ 的值。

(2) 神經網路（Neural Network, NN）

我們使用不同參數實驗，最終選擇P@1最佳時的參數作為最後使用之參數。對於C+ 則和邏輯迴歸相同設定為5。

表八、實驗結果（節錄）

參數選擇				結果
#filters	filter size	dropout prob.	learning rate	P@1
700	6	0.4	0.0001	0.846516
500	6	0.4	0.0001	0.840335
300	8	0.4	0.0001	0.821582
300	6	0.4	0.0001	0.807543
300	10	0.6	0.0001	0.789209

表九、最終使用參數

參數	最終使用
#filters	700
filter size	6
dropout prob.	0.4
learning rate	0.0001

P@1表現最佳為0.846516，我們選擇達到此結果的參數作為最終使用之參數。

六、預測

在疾病預測系統中，通常針對每個患者，向醫生展示少數疾病及描述。考慮以下的多標籤系統，對於每個病人，預測出的前面K個疾病顯示出給醫生，並且K是一個小值（例如：5, 10）。接著我們在預測階段提出以下的設置。

(一) 對於每個病人，從第一階段訓練出的模型獲得以下兩點
1. 排名後的常見疾病
2. 排名後的罕見疾病

(二) 來自第二階段訓練的二元成本導向模型預測病人是否具有罕見疾病，也就是是否與少標籤有關係

1. 有關，前1~K-1個位子放前面1~K-1個常見疾病，位子K放排名最前面的罕見疾病
2. 無關，前1~K個位子均放常見疾病

二元成本導向問題允許我們調整具有罕見疾病的病人，並保留前K個位子中的最後一個位子給排名最前面的罕見疾病。

七、比較

我們將上述的方法與另外兩個設定比較。

(一) 不具有第二階段，只經過第一階段的多標籤分類，直接將排名最前面的疾病顯示給使用者。此設定也是現行使用多標籤分類在疾病分類的作法。

(二) 將預測的前K個疾病中，第K個位子完全保留給罕見疾病。這樣的設定是我們提出的兩階段預測方式的一個極端特例。若我們把第二階段中罕見疾病的權重設為遠比非罕見疾病高，則所有病人都會被判斷為具有罕見疾病，這就相當於將第K個位子完全留給罕見疾病。

八、製作系統

利用Matlab製作疾病預測系統。

(一) 選取病人病歷文字檔。

(二) 顯示病人病歷內容。

(三) 預測疾病類別 ICD編號與疾病名稱。

肆、研究結果

一、實驗設定

(一) 將MIMIC-III資料分成Training Set與Test Set。

表十、Training Set與Test Set的總數與Few-shot數

	# of Instances	# of Instances with Few-shot	# of Labels
Training	49,354	6,824	8,758
Test	3,372	600	8,758

我們定義在資料集中出現少於5次的標籤為少標籤。由上表可知，病人總數在Training Set和Test Set中分別是49,354與3,372，而其中患有罕見疾病的病人數分別是6,824與600。標籤（Label）總數，即疾病總數為8,758。

(二) 訓練與檢驗

在訓練後，以Test Set進行檢驗，檢驗顯示給使用者前5個疾病時的表現。

1. 第一階段與第二階段

表十一、不同階段表示與涵義

表示	Multi-label method	Binary cost-sensitive method
涵義	即第一階段使用之疾病分類的方法。	第二階段二元成本導向判斷病人是否具有罕見疾病使用之方法。

第一階段使用的方法分別為LR與NN，而第二階段分別為不作第二階段處理（N/A）、對於所有病人均保留地K位子給罕見疾病（All），及二元成本導向決定是否保留第K位給罕見疾病，其中又分為LR與NN。

表十二、第一階段表示與涵義

表示	N/A	All	LR	NN
涵義	現有方法，未進行第二階段處理	前K位的第K位保留給罕見疾病	邏輯迴歸	神經網路

2. 結果表示

表十三、評估方法表示

表示	P@K	R@K	PSP@K	#few-shot predicted in Kth position	#correct few-shot predictions
涵義	Precision精確率	Recall召回率	Propensity Weighted Precision	第K個位置預測為罕見疾病的病人數量	預測結果前K個位置含有罕見疾病且預測正確的病人數

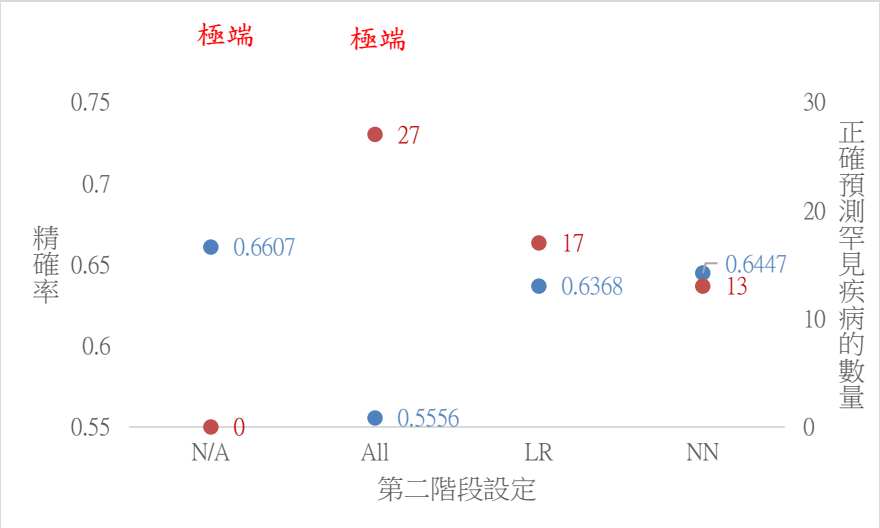
我們考慮k=5。

二、整體表現

(一) 第一階段使用LR

表十四、使用LR時的預測表現

Multi-label method	Binary cost-sensitive method	P@5	R@5	PSP@5	#few-shot predicted in Kth position	#correct few-shot predictions
LR	N/A	0.6607	0.2226	1.2485	0	0
	All	0.5556	0.1895	1.0505	3,372	27
	LR	0.6368	0.2150	1.1929	1,003	17
	NN	0.6447	0.2179	1.2125	620	13



圖十一、第一階段使用LR時第二階段不同設定的結果（藍色的點為精確率，紅色的點為正確預測罕見疾病的數量）前兩個第二階段設定造成極端的結果：精確率與罕見疾病的預測分別大好或大壞

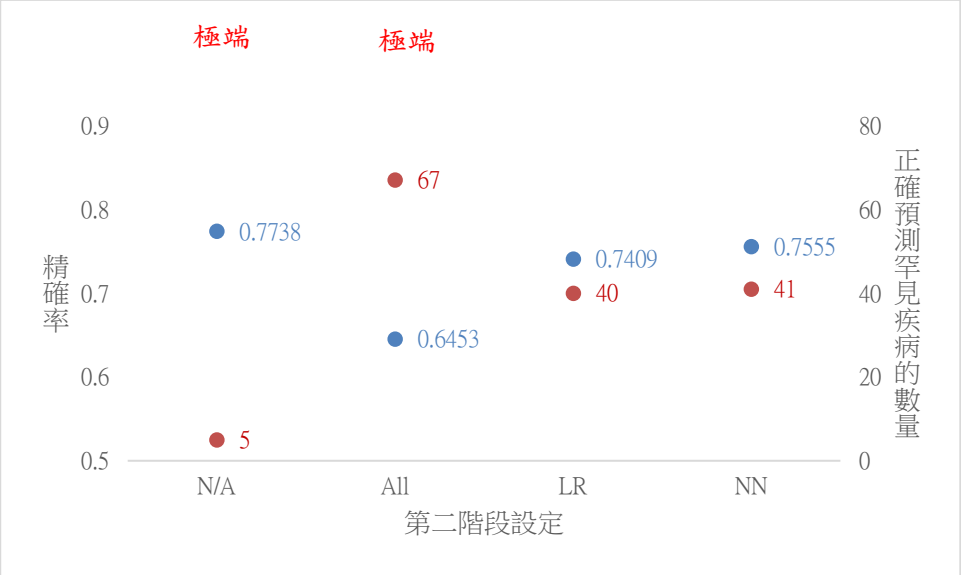
上面表格為當第一階段使用LR，第二階段不同的設定分別得到不同的結果。不進行第二階段時，精確率為0.6607，召回率為0.2226，PSP為1.2485，而未能預測任何病人有罕見疾病。將所有病人的第5個位置留給罕見疾病時，精確率為0.5556，召回率為0.1895，PSP為1.0505，而所有病人的第5個位置均預測罕見疾病，而有27個病人正確地被預測有罕見疾病。若第二階段使用LR進行二元成本判斷病人是否有罕見疾病，則精確率為0.6368，召回率為0.2150，PSP為1.1929，並且1,003位病人的第5個位置預測為罕見疾病，正確的有17個。第二階段使用NN則精確率為0.6447，召回率為0.2179，PSP為1.2125，其中620位病人在第5個位置預測罕見疾病，正確的有13個。

可以看出精確率方面未經任何處理的表現最好，但未能正確預測罕見疾病。而將第5個位置都留給罕見疾病時，正確預測罕見疾病的數量則最高，精確率卻下降超過10%。經過第二階段判斷病人是否具有罕見疾病後不論是使用LR或是NN的結果則均是正確預測出罕見疾病的數量在前面兩者之間，而準確率下降均小於3%。

(二) 第一階段使用NN

表十五、使用NN時的預測表現

Multi-label method	Binary cost-sensitive method	P@5	R@5	PSP@5	#few-shot predicted in Kth position	#correct few-shot predictions
NN	N/A	0.7738	0.2616	1.6081	11	5
	All	0.6453	0.2210	1.3710	3,372	67
	LR	0.7409	0.2523	1.5468	1,003	40
	NN	0.7555	0.2565	1.5760	620	41

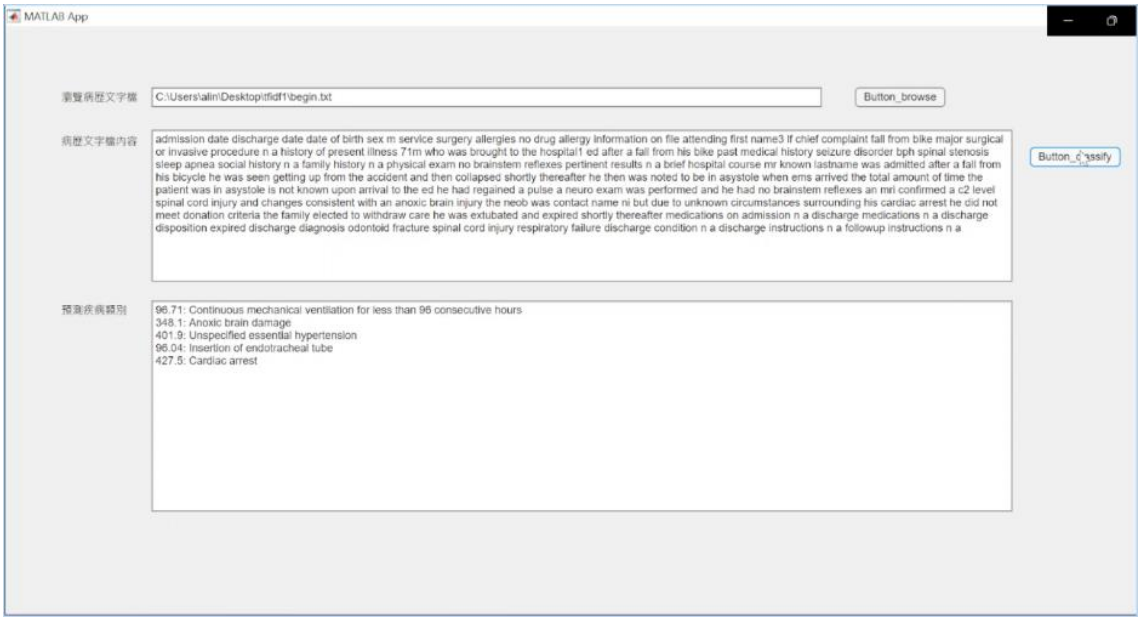


圖十二、第一階段NN時第二階段不同設定的結果（藍色的點為精確率，紅色的點為正確預測罕見疾病的數量）前兩個第二階段設定造成極端的結果：精確率與罕見疾病的預測分別大好或大壞

上面表格為當第一階段使用NN，第二階段不同的設定分別得到不同的結果。不進行第二階段處理的精確率達77.3%，表現最好，但正確預測出病人罕見疾病的數量只有5個。而第5個位置均放置罕見疾病時精確率較前者下降超過10%，但正確預測病人罕見疾病的數量達67個。若第二階段使用LR進行二元成本判斷病人是否有罕見疾病，精確率為0.7409，有1,003位病人的第5個位置預測為罕見疾病，正確的有40個。第二階段使用NN則精確率為0.7555，其中620位病人在第5個位置預測罕見疾病，正確的有41個，而經過第二階段決定是否將第5個位置留給罕見疾病這樣的設定下，精確率低於未經處理的約2-3%。

四、製作疾病預測系統

我們成功地提出能兼顧罕見疾病預測與整體表現的演算法，並運用Matlab實作系統，可藉由輸入病人病歷文字檔預測病人可能患有的疾病及編碼。



圖十三、疾病預測系統

伍、討論

一、經過第二階段與否結果討論

(一) 未經任何處理

1. 整體表現最佳但罕見疾病預測表現差

在上述的結果中我們可以看出未經任何第二階段處理，直接進行現有的方法雖然有較高的精確率，但在正確預測出病人罕見疾病的方面表現不理想，每個設定下均正確預測出最少的數量。在第一階段使用LR的方法中，未能正確預測出任何的罕見疾病。

由這樣的結果可以知道若沒有做任何的調整，直接進行疾病的分類，罕見疾病幾乎不會被顯示給使用者。罕見疾病在訓練過程中出現次數少，相較於出現次數多的疾病不會對整體結果造成很大的影響，因此訓練過程中雖然提升準確率，只能代表對準確率影響較大的常見疾病的預測進步，罕見疾病的預測並未提升。也就是說在多標籤分類中，少標籤容易被忽略。

(二)第K位一律留給罕見疾病

1. 正確的預測有罕見疾病的病人的罕見疾病數量最多。當我們將第K個位子均留給罕見疾病，則被成功預測出的罕見疾病數量會是我們提出的方法所能得到最好的結果。
2. 精確率、召回率、PSP顯著下降

由結果可以看出，在這樣的設定下，準確率、召回率顯著下降。由於病人總數為3,372，而患有罕見疾病的病人數為600，只佔全部的17.8%，所以當對每位病人的預測中的第K個位置均留給罕見疾病，大部分在這個位置的疾病的預測都是錯誤的。

所以這樣的設定雖然能正確找出較多具有罕見疾病的病人的罕見疾病，但會大幅降低整體的表現。這樣的結果指出加強罕見疾病預測的同時，需要考慮預測罕見疾病所帶來的成本。

(三) 經過第二階段的判斷決定是否將第K位留給罕見疾病

1. 正確的預測病人的罕見疾病數量高於未經處理時

我們提出的方法在第二階段判斷病人是否患有罕見疾病，若患有罕見疾病，才將第K個位子留給罕見疾病。經過第二階段後，第K位放置罕見疾病的病人數量相對小，且正確預測出病人罕見疾病的數量相當不錯。不會像第一種方法幾乎沒有預測出任何罕見疾病，也不會像第二種方法對任何病人在第K個位置均放罕見疾病。

例如第一階段使用NN時，若以現有方法，也就是不進行任何的第二階段直接預測，正確預測出病人罕見疾病的數量只有5，但我們所提出的方法在第二階段使用NN時，正確預測出病人罕見疾病數量為40，為現有方法的8倍。

2. 整體表現與未經處理時差距小

很重要的一點是，在這樣的設定下，整體的表現雖然仍然不如未經過第二階段處理時的表現，但並未造成太大的影響。

此方法在第K個位置放置罕見疾病的病人數量相對小，這樣能避免整體表現的下降。若在第二階段將所有病人的第K個位置都留給罕見疾病，因患有罕見疾病的病人占少數，這樣的設定導致大多數病人的第K個位置預測的疾病錯誤，而整體表現下降。而我們提出的方法在經過第二階段判斷後，第K個位置

放置罕見疾病的病人數量相對小，因此不再遇到沒有患有罕見疾病的病人第K位卻預測為罕見疾病的問題，第K位預測錯誤的數量下降，整體表現也不會有太大的影響。

也就是說，因為我們進行第二階段時先判斷病人是否具有罕見疾病，而能夠調整第K個位置留給罕見疾病的病人數量，以達成改善罕見疾病的預測與維持整體準確率的平衡。

改良方法 正確預測病人罕見疾病的數量能達現有方法的8倍，且精確率只下降約0.02。

表十六、經第二階段與否的預測表現比較表

經第二階段與否 預測結果	未經任何處理	第K位一律留給罕見疾病	經過第二階段的判斷決定是否將第K位留給罕見疾病
整體表現	最佳	最差	次之
罕見疾病	最差	最佳	次之
是否能達到平衡	極端	極端	最佳

二、疾病預測模型訓練結果討論

整體而言，不論在第一階段或第二階段，NN均略優於LR。這兩個階段使用的機器學習方法是可以完全獨立的。我們的實驗結果大略指出在每個階段我們必須使用在該階段最好的機器學習方法，組合後能夠達到最好的整體表現。

三、未來展望

(一) Zero-shot 的預測

有些疾病在訓練階段出現次數為零，稱為zero-shot，但它有可能在使用時出現，因此仍有重要性。若少標籤的預測能有改善，或許在zero-shot的預測上也能有突破。

(二) 使用此概念在不同領域

將多標籤分類分為兩階段再進行預測已達到少標籤預測的改善，這樣的概念不只能運用在疾病分析上，也能使用在其他多標籤分類問題上。文章分類、

網路上留言分類等都能使用，我們期盼能藉由我們所提出的方法能改善各類多標籤分類問題中少標籤的預測。

陸、結論

在此篇作品中，我們提出將疾病的多標籤分類切分成兩個階段，其中第一階段使用現有的多標籤分類方法，而第二階段利用二元成本導向判斷病人是否具有罕見疾病，再根據第二階段的結果決定最後預測病人可能患有的疾病中，是否將最後一個預測位置保留給罕見疾病。我們的結果指出兩階段的設計能夠得到同時兼顧罕見疾病預測與整體表現的最佳平衡。

這樣的方法不會過度專注在少標籤，亦不會完全忽略少標籤。能夠在避免少標籤完全不被預測的同時亦避免疾病預測整體的表現下降過多，能使用在真實情況，具有高的實用性，並且已成功實作疾病預測系統。

柒、參考資料及其他

[1] Rios, A., & Kavuluru, R. (2018). Few-shot and zero-shot multi-label learning for structured label spaces. Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence (IJCAI), 2847-2853. https://aclanthology.org/D18-1352/

[2] Mullenbach, J., Wiegreffe, S., Duke, J., Sun, J. & Eisenstein, J. (2018). Explainable prediction of medical codes from clinical text. Proceedings of the Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL), 1101-1111. https://arxiv.org/abs/1802.05695

[3] Wu, G., Tian, Y.J., & Liu, D. (2018). Cost-sensitive multi-label learning with positive and negative label pairwise correlations. Neural networks, 108: 411-423. https://pubmed.ncbi.nlm.nih.gov/30312958/

[4] Fan, R.E., & Lin, C.J. (2007). A study on threshold selection for multi-label classification. Technical report, Department of Computer Science, National Taiwan University. https://www.semanticscholar.org/paper/A-Study-on-Threshold-Selection-for-Multi-label-Fan-Lin/f3ebf945aba8d70b8d7daf14021fe1220752f0f7

[5] Chalkidis, I., Fergadiotis, M., Kotitsas, S., Malakasiotis, P., Aletras, N., & Androutsopoulos, I. (2020). An empirical study on largescale multi-label text classification including few and zero-shot labels. Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), 7503-7515. https://arxiv.org/abs/2010.01653

[6] Khandagale, S., Xiao, H., & Babbar. R. (2020). Bonsai: diverse and shallow trees for extreme multi label classification. Machine Learning, 109: 2099-2119. https://arxiv.org/abs/1904.08249

[7] Liu, J., Yang, T., Chen, S., & Lin, C. (2021). Parameter Selection: Why We Should Pay More Atte ntion to It. Proceedings of the 59th Annual Meeting of the Association of Computational Linguistics (ACL). https://arxiv.org/abs/2107.05393

[8] Bhatia, K., and Dahiya, K., and Jain, H., and Kar, P., and Mittal, A., and Prabhu, Y., & Varma, M. (2016). The extreme classification repository: Multi-label datasets and code. http://manikvarma.org/ downloads/XC/XMLRepository.html

【評語】190021

1. 此作品提出將疾病的多標籤分類切分成兩個階段，透過現有的多標籤分類方法加上二元成本導向方法，判斷病人是否具有罕見疾病。題目具有實用價值，且可處理少標籤之模型訓練和預測的問題，是一完整的作品。

2. 建議未來可以採取更新的資料集進行實驗，且對於不同的模型也可進行實驗，以進行比較分析。