**AI CUP 2023 春季賽**

**多模態病理嗓音分類競賽報告**

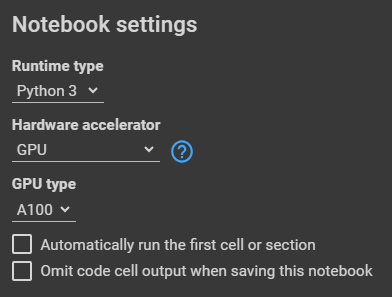
隊伍：TEAM\_3687

隊員：黃永璿（隊長）、陳佳辰、林介元、吳芮妍、楊于璇

Private leaderboard：分數 0.5796/ Rank 19

**壹、環境**

本團隊於本次競賽中使用的是Google Colaboratory（Colab），有關於音檔的硬體加速器改為GPU（如圖一），文字檔使用CPU，作業系統為Ubuntu Linux，程式語言為Python 3（v3.7.13），Colab中，因此使用的函示庫與環境均為Colab原始設定。



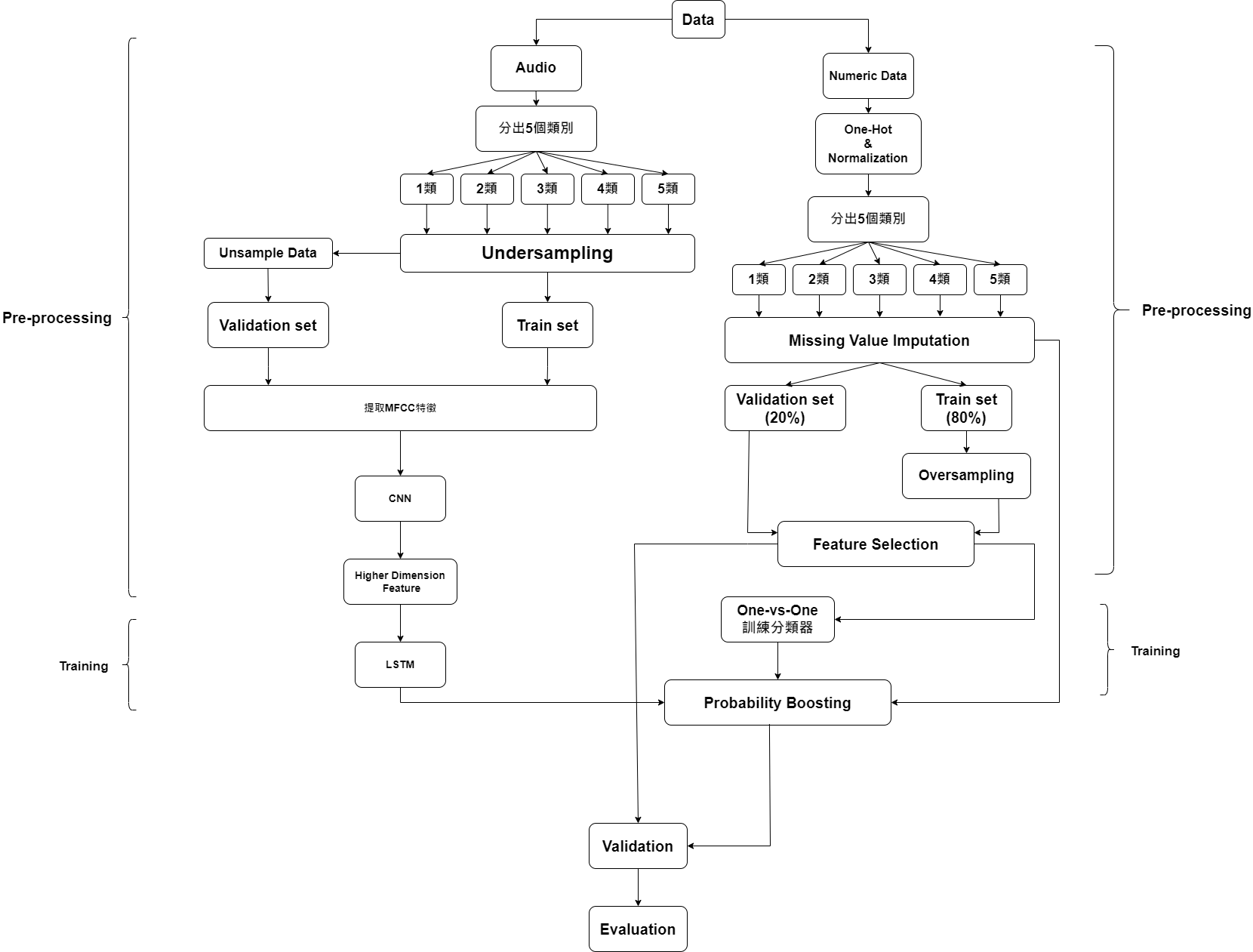
圖一、Colab 設定

**貳、演算方法與模型架構**

本團隊在次競賽中發現了最關鍵的兩個問題是：多類別分類(Multi-class Classification)與資料不平衡（Class Imbalanced）。

現今有許多優秀的演算法在二元分類的問題上可以有很好的成效，然而本競賽為一個多類別分類的問題，在二元分類問題上表現良好的演算法不一定能很好地解決多類別分類的問題；另一方面資料集的不平衡會導致在訓練模型時，對於多數的類別會過度學習進而產生預測偏好，對於少數的類別則是沒有足夠多的資料可以學習。

為了解決上述兩個問題，本團隊對不同模態的資料分別進行了不同的處理，圖二為本團隊在本次競賽中所使用的模型架構。



圖二、實驗模型架構

對於數值型資料，本團隊使用了One-Versus-One(OVO)的架構(Peter Mills, 2021)，OVO架構將多類別分類的問題轉變成多個二分類的問題，最後再將每個二元分類器的結果依權重進行投票得到一個結果。

而對於音訊型態的資料，本團隊在實驗過程中發現資料集不平衡的問題

對於訓練模型的成效影響重大，因此在前處理的部分先將完整資料集依照5個類別進行切分，然後將所有類別的資料比數做undersampling至與比數最少的類別一樣(在本實驗中為第5類的32筆)，得到一個平衡的資料集作為訓練資料集，而不在訓練資料集裡的資料則做為測試與驗證資料集。完成資料切分後提取MFCC特徵獲得資料的特徵向量，由於訓練資料集太小，因此本團隊再進一步使用CNN提取更高維的特徵後使用LSTM作為訓練分類器的模型。

最後將音訊資料的分類器與數值型資料的分類器進行融合，得出最終的分類結果。

**參、創新性**

本團隊於本次競賽中之創新性主要展現於兩個階段:前置資料處理流程(Pre-processing)及訓練方式流程(Training)，以下依據圖二進行說明。首先，在前置資料處理流程(Pre-processing)階段，依序完成Data Transformation、One-hot encoding(處理雜訊值)、Normalization(採用Min-max scaling)**、**Data Splitting(依據Class切分)、Missing Value Imputattion(採用KNN Imputattion)等步驟。在此階段中，最大的創新性在於補值的部分：本團隊先將所有資料依照Class切分，接著再個別針對各個Class進行補值，如此一來相同的Class所得補值會較集中趨近。

此外，在訓練方式流程(Training)階段，依序完成Train set/Validation set切分、Oversampling處理資料不平衡問題、Feature Selection選擇出有效特徵、One-vs-One方法訓練分類器以進行多類別預測、及採用Probrablitiy Boosting方法來訓練並優化模型。在此階段中，最大的創新性在於兩個部分。第一，採用One-vs-One方法解決多分類問題。相較於 One-vs-Rest 方法由於樣本數量可能的偏向性所帶來的不穩定性，One-vs-One方法是一種相對穩健的擴展方法。第二，採用Probrablitiy Boosting方法訓練並優化模型。本團隊採用的Probrablitiy方法有別於傳統Voting solution採用「一分類器投一票，票票等值」之概念，進而改採加總所有分類器以「Probrablitiy機率」輸出之機率值，作為分類物件屬於什麼類別的判斷依據。而透過各個分類器以機率輸出結果的過程，以及經過模型多次「Boosting」的迭代訓練(即結合Probrablitiy和Boosting方法)，此模型的訓練學習過程目的就是要最小化預期風險，將其中損失函數的計算轉為計算機率密度函數。本團隊將此創新方法實作於本次競賽之多任務分類任務。

另一方面，於音訊分類模型的訓練中，有別於競賽提供之參考Baseline模型中僅用一般常見之多層感知器（Multilayer Perceptron, MLP）(Pal et al.,1992)訓練模型，本團隊首先採用卷積神經網絡(Convolutional Neural Network, CNN)(O'Shea and Nash, 2015)提取特徵，再採用LSTM訓練模型。這是由於CNN在音訊處理中具有捕捉局部特徵等優點(Liu et al., 2017)。此外，經過CNN模型提取出之特徵再投入LSTM模型，其模型表現比單獨採用CNN或單獨採用LSTM訓練模型結果更好。因此，使其成為本團隊處理音訊數據的創新模型架構。

**肆、資料處理**

根據此架構模型的訓練方式（圖二），可以區分成兩個部分，1. 數值資料處理2. 音訊資料處理。

1. 數值資料處理

* Data Transformation

針對Sex、Noise at work、PPD此3個欄位處理，以下依照各欄位說明

* + Sex

針對此欄位做Label encoding，將原先1變為0（男性）、2變為1（女性）

* + Noise at work

針對此欄位做Label encoding，將原先1變為0（否）、2 變為1 （有一點）、3變為2（很吵）

* + PPD

針對同一列Smoking為0的情況下（從未）， 將PPD設為0（一天0包菸）

* Normalization

針對Disease category、ID以外欄位做**Min-max scaling**，將其他欄位正規化

* Data Splitting(By Class)

將資料依照Disease category類別切成5份資料（Class1-Class5）

* Missing Value Imputation

將先前依照Disease category類別切分的資料個別做**KNN Imputattion**

2. 音訊資料處理

* Data Splitting (By Class)

將資料依照Disease category類別切成5份資料（Class1-Class5）

* Undersampling

根據資料比數最少的Disease category class(Class 5)來決定各Class Under sampling數量（各 Class 32筆）作為train set；其餘沒被sample的資料作為validation set

* MFCC特徵萃取

使用**Librosa**來取得個音檔的梅爾頻率倒譜係數（Mel-Frequency Cepstral Coefficients）

* Higher Dimension Feature

使用**CNN**來產生更多features

**伍、訓練方式**

根據此架構模型的訓練方式，可以區分成兩個部分，(1)數值分類模型的訓練 (2) 音訊分類模型的訓練。

1. 數值分類模型的訓練

* Train set/ Validation set 切分

首先，在數值資料部分，經過前述的前置資料處理流程(Pre-processing)之後，我們可以獲得5個類別且完整無缺失值的資料集。之後，我們會分別從這五個類別中以8:2的比例隨機抽取資料作為訓練資料集(train set)和測試資料集(validation set)並輸出成csv檔進行保存，這麼做是因為原始資料集是一個極度不平衡的資料集(最大類別有537筆資料，最小類別只有33筆資料)，為了避免因為隨機抽取資料集時，小類別資料未被選中而被排除在train set外的可能性。同時，為了提供後續實驗比較的基礎，以避免每次執行都隨機抽取不同的資料作為訓練資料。

* Oversampling

有了train set之後，我們會分別使用SMOTE[(Chawla et al., 2002)](https://www.zotero.org/google-docs/?a90JtM)、ADASYN[(Haibo He et al., 2008)](https://www.zotero.org/google-docs/?74E3IF)來做oversampling 以降低資料不平對分類器學習的負面影響。此處我們決定將train set所有類別都擴增至500筆資料，train set共計2500筆資料。

* Feature Selection

在數值資料當中有多個特徵欄位，為了讓分類器學習重要的特徵，我們使用Pearson、Spearman、Kendall三種相關係數來分析各個特徵對Disease Category的相關性。並使用不同的挑選標準來得出最佳的特徵組合。

**陸、分析與結論**

對於數值型態的資料處理上，經團隊實驗後發現不同補值的方法，差異不顯著，本團隊認為相較於缺失值的問題，更關鍵的是多分類任務以及資料集不平衡。而比起OVO架構，probrablitiy regression的方式來得更有效，因此可以得出結論：應該為不同的分類器給予權重值，效果會比較好，因為可以避免單個或少數個分類器的偏誤。

而音訊資料的部分，分類器的效能有待加強，在未來可以嘗試使用resnet做特徵提取。

**柒、程式碼**

* 文字處理

[Text Data - Google 雲端硬碟](https://drive.google.com/drive/folders/10UT4B1q-HatqFUnwQC7IQlvjr7uSYUZg)

* 音訊處理

[Audio\_CNN+LSTM.ipynb](https://colab.research.google.com/drive/1Vx4SNfI6SZQgqCARxiv2EsCGlbTd1TOY?usp=sharing)

**捌、使用的外部資源與參考文獻**

[Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research*, *16*, 321–357. https://doi.org/10.1613/jair.953](https://www.zotero.org/google-docs/?7w30M7)

[Haibo He, Yang Bai, Garcia, E. A., & Shutao Li. (2008). ADASYN: Adaptive synthetic sampling approach for imbalanced learning. *2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence)*, 1322–1328.](https://www.zotero.org/google-docs/?7w30M7) <https://doi.org/10.1109/IJCNN.2008.4633969>

Liu, W., Wang, Z., Liu, X., Zeng, N., Liu, Y., & Alsaadi, F. E. (2017). A survey of deep neural network architectures and their applications. *Neurocomputing*, *234*, 11-26.

O'Shea, K., & Nash, R. (2015). An introduction to convolutional neural networks. *arXiv preprint arXiv:1511.08458*.

Pal, S. K., & Mitra, S. (1992). Multilayer perceptron, fuzzy sets, and classification. *IEEE Transactions on neural networks*, *3*(5), 683-697.

Peter Mills (2021). Solving for multi-class: a survey and synthesis