Analysis using ML, DL

Group 3

尹子維、李承祐、黃亮臻、張立勳

2024-06-11

目標

預測五個變數 (Bone, Brain, Kidney, Liver, Lung) 是否有轉移。

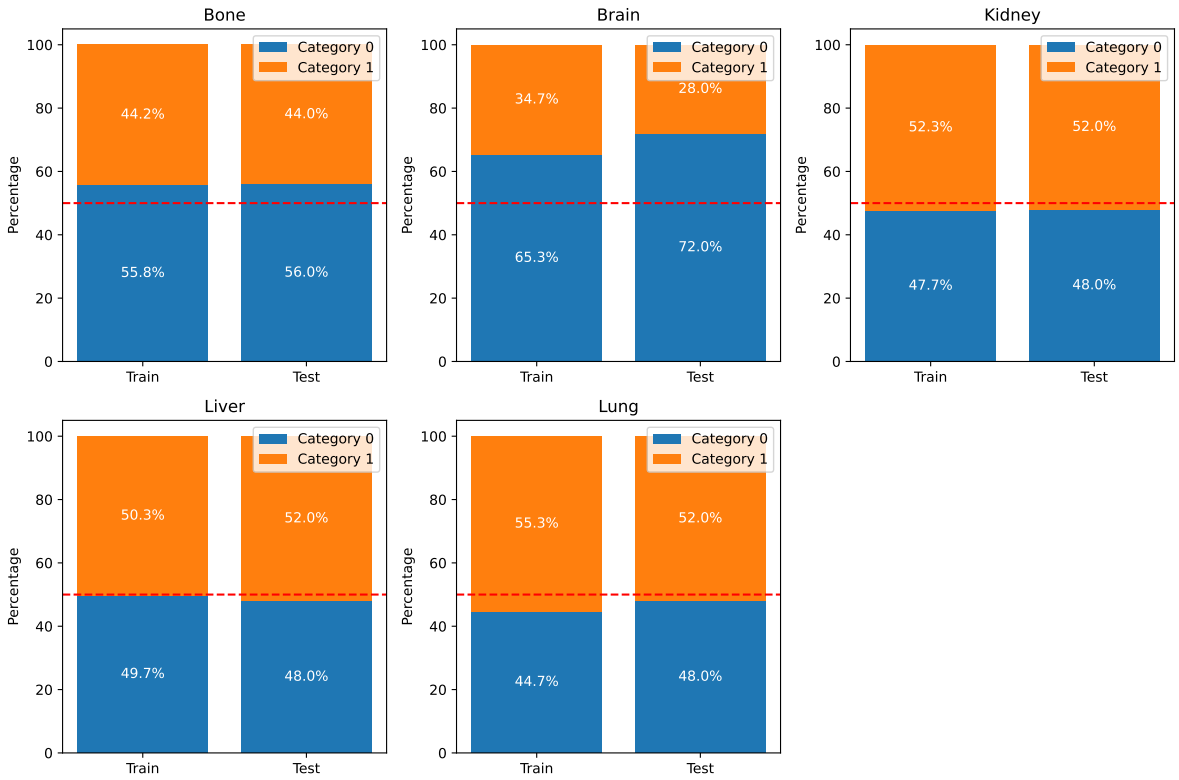
方法

- 深度學習: Multi-head Neural Network
- 機器學習: Decision Tree, Random Forest, LightGBM, CatBoost

預處理

劃分資料集

將資料按照 9:1 切分訓練集與測試集。針對五個目標變數的轉移狀態，此資料共有 28 種可能的狀態組合。採用分層切分的方式，確保各個組合在訓練集與測試集的比例盡量保持一致。



由上圖發現，`Brain` 存在些許資料不平衡問題 (1.89 倍)，其餘變數則不是很明顯，後續分析將嘗試針對 `Brain` 變數做上採樣 (SMOTE)。

Multi-head Neural Network

目標

希望做到一個模型同時預測多個目標變數。

模型設計

共享層

包含一個全連接層，配合 ReLU 激活函數，其主要作用是提取輸入資料中的通用特徵。這層的輸出維度設定為 128 維。

五個獨立的輸出層

在共享層之後，架構分出五個獨立的輸出層，每個輸出層都由一個全連接層構成。每個輸出層都專門負責預測一個特定的目標變數。這種設計允許模型對每個目標進行專門的學習和預測，同時基於共享層的特徵，加強模型對各目標之間可能存在的隱含關聯的理解。

損失函數

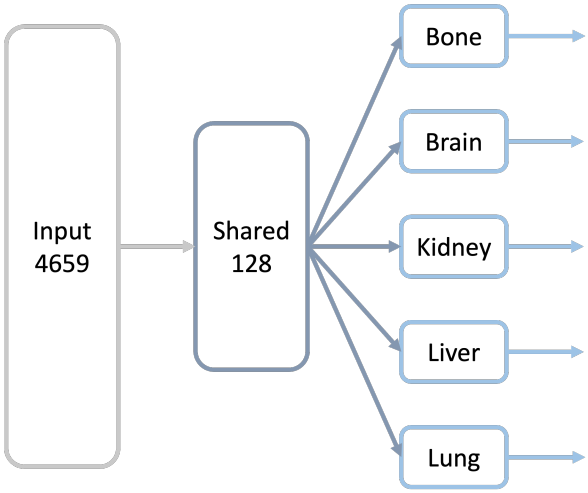
採用二元交叉熵（Binary Cross-Entropy）作為損失函數，對每個獨立輸出層的預測結果計算損失。由於這是一個多目標的預測任務，模型會計算所有輸出層的損失總和，以此來進行梯度下降並更新網路的權重。

優化器

Adam

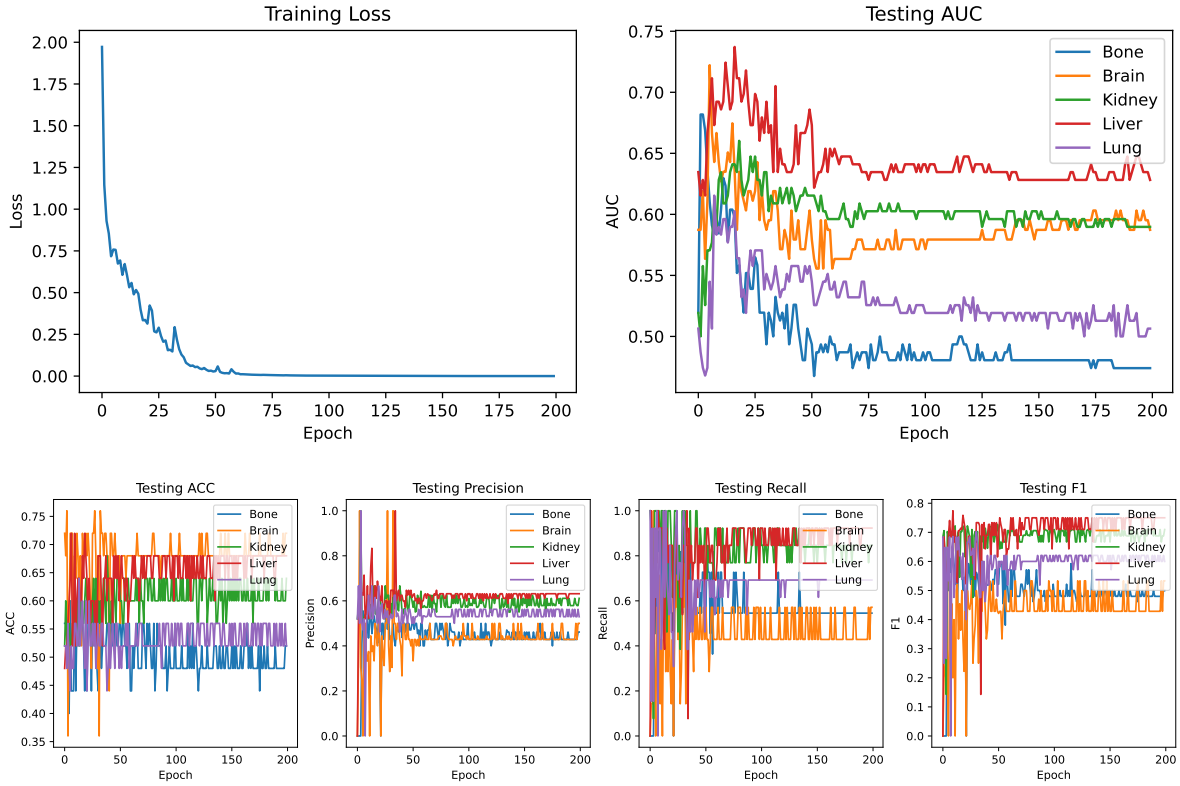
學習率

0.0005



共有 $(4659 + 1) \times 128 + (128 + 1) \times 5 = 597125$ 個參數待估計。

模型表現



深度學習模型對隨機初始值特別敏感，尤其當樣本數較少時，這一點更為明顯。在此次實驗的五個目標變數的預測中，AUC 值的波動範圍從 0.4 到 0.7 不等。但若是初始值設定不佳，模型有時會在初期傾向將所有樣本預測為全正或全負，導致預測結果極不穩定。因此，在樣本量有限的情況下，依賴深度學習可能不是理想的選擇。

Machine Learning Method

目標

為目標變數: Bone, Brain, Kidney, Liver, Lung 分別建立五個模型，並觀察對這五個目標最有影響的變數。

方法

由於希望考慮交互作用項，以及後續方便解釋特徵重要性，這裡皆使用 Tree-based 模型。

- Decision Tree
- Random Forest
- LightGBM
- CatBoost

Decision Tree

	Accuracy	Precision	Recall	F1	AUC
Bone	0.48	0.3750	0.2727	0.3158	0.4578
Brain	0.52	0.2222	0.2857	0.2500	0.4484
Kidney	0.32	0.3750	0.4615	0.4138	0.3141
Liver	0.60	0.6154	0.6154	0.6154	0.5994
Lung	0.52	0.5385	0.5385	0.5385	0.5192

Random Forest

	Accuracy	Precision	Recall	F1	AUC
Bone	0.64	0.6250	0.4545	0.5263	0.6656
Brain	0.68	0.3333	0.1429	0.2000	0.4206
Kidney	0.64	0.6111	0.8462	0.7097	0.7115
Liver	0.52	0.5333	0.6154	0.5714	0.6538
Lung	0.40	0.4444	0.6154	0.5161	0.3622

LightGBM

	Accuracy	Precision	Recall	F1	AUC
Bone	0.56	0.5000	0.6364	0.5600	0.5455
Brain	0.48	0.1250	0.1429	0.1333	0.4206
Kidney	0.52	0.5238	0.8462	0.6471	0.5577
Liver	0.60	0.6154	0.6154	0.6154	0.5833
Lung	0.48	0.5000	0.7692	0.6061	0.5577

CatBoost

	Accuracy	Precision	Recall	F1	AUC
Bone	0.56	0.5000	0.5455	0.5217	0.6299
Brain	0.64	0.3333	0.2857	0.3077	0.4921
Kidney	0.48	0.5000	0.8462	0.6286	0.5897
Liver	0.52	0.5333	0.6154	0.5714	0.5897
Lung	0.48	0.5000	0.7692	0.6061	0.5641

SMOTE

僅針對 `Brain` 變數做上採樣，將少類的樣本補到多數類的 80%。

```
Brain
原資料: 轉移 130, 沒轉移 69
SMOTE後: 轉移 130, 沒轉移 104
```

Model	Accuracy	Precision	Recall	F1	AUC
DecisionTree	0.48	0.3750	0.2727	0.3158	0.4578
RandomForest	0.72	0.8333	0.4545	0.5882	0.7695
LightGBM	0.60	0.5714	0.3636	0.4444	0.6558
CatBoost	0.68	0.7143	0.4545	0.5556	0.6234

將做過 SMOTE 上採樣的資料進行訓練，所有模型的表現皆提高了。除了 DecisionTree 只有微幅上升以外，其餘模型 RandomForest, LightGBM, CatBoost 皆提高了 0.2 以上。

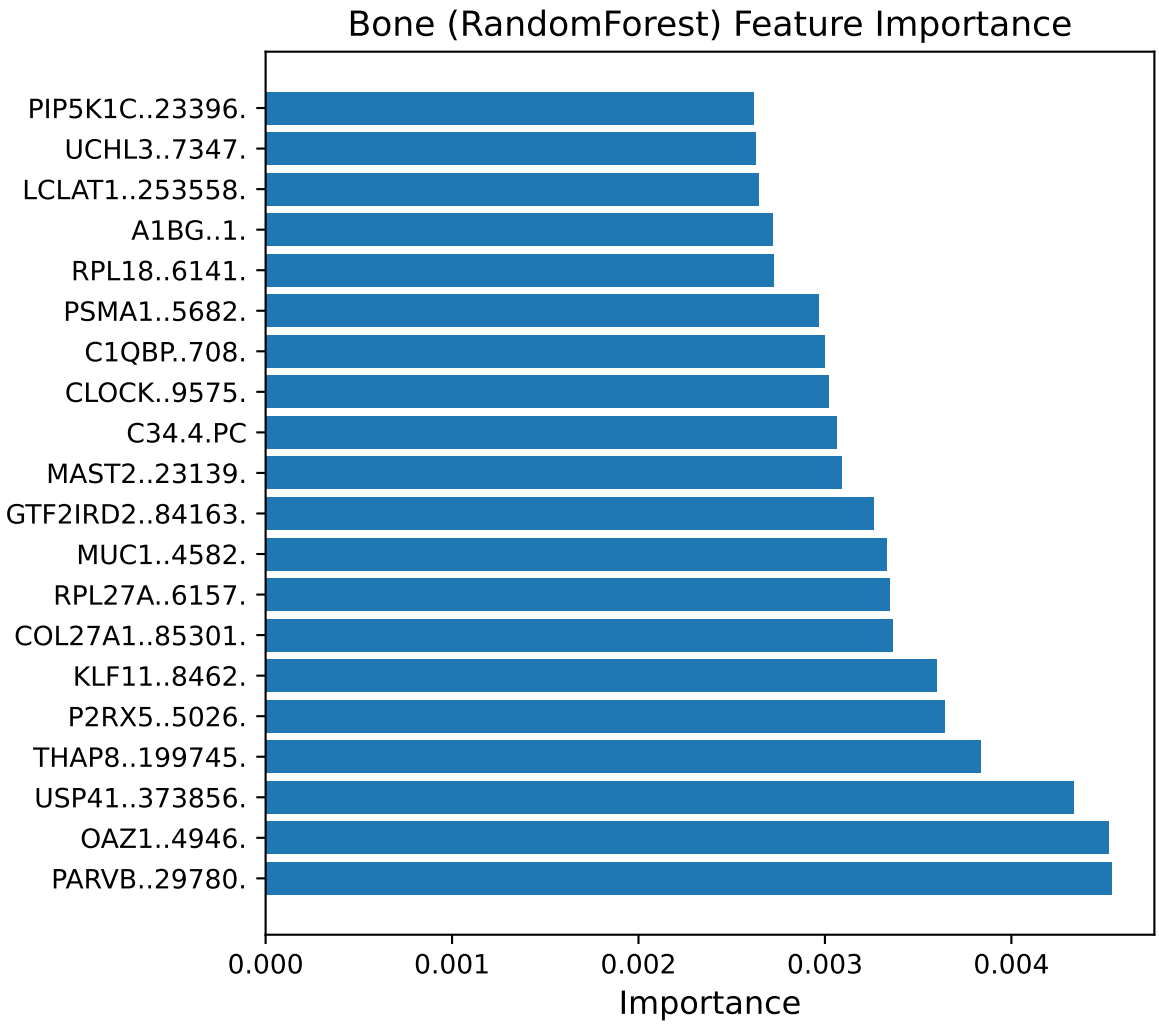
各個目標變數對應 **AUC** 最高之模型

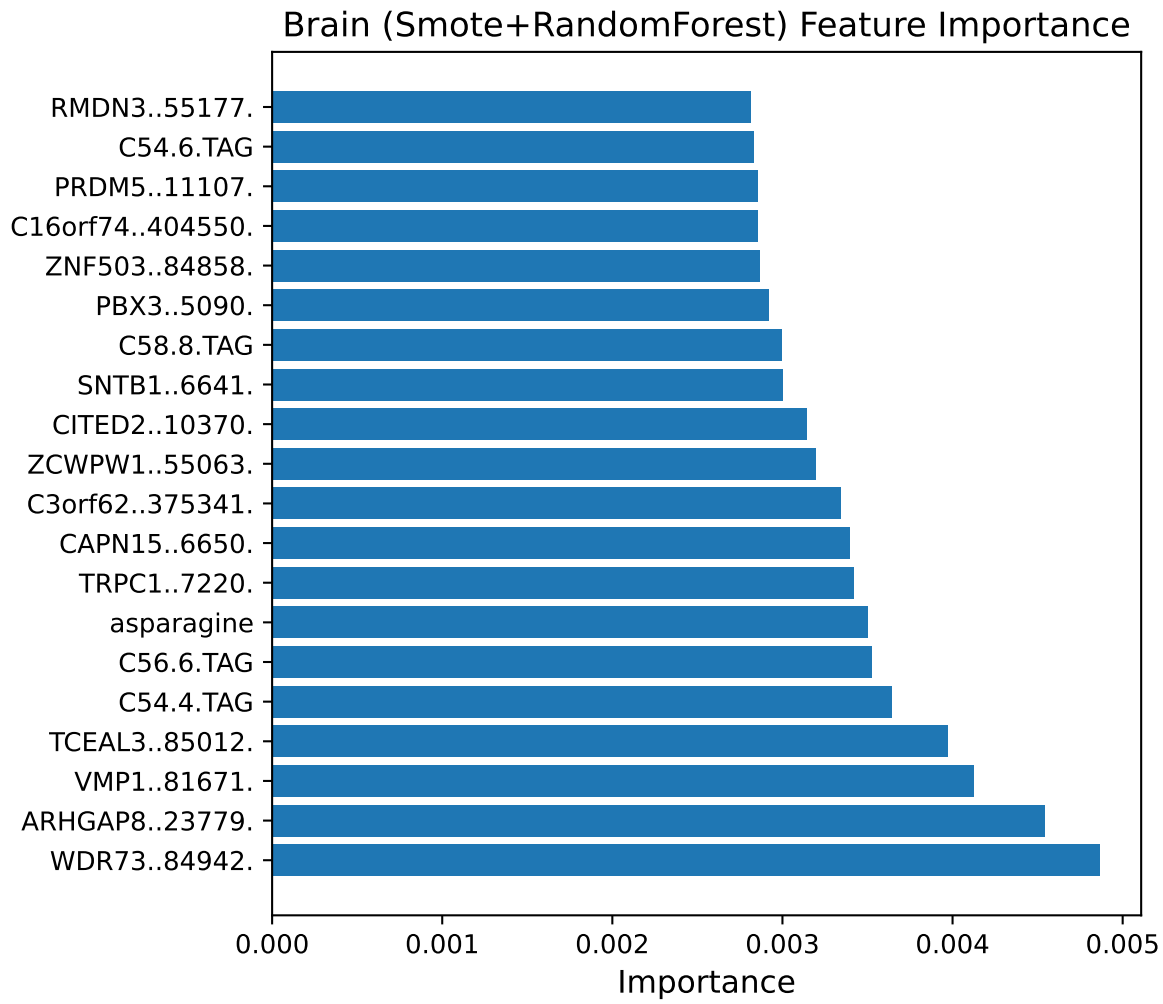
	Bone	Brain	Kidney	Liver	Lung
Accuracy	0.64	0.72	0.64	0.52	0.48
Precision	0.625	0.8333	0.6111	0.5333	0.5
Recall	0.4545	0.4545	0.8462	0.6154	0.7692
F1	0.5263	0.5882	0.7097	0.5714	0.6061
AUC	0.6656	0.7695	0.7115	0.6538	0.5641
Source	RandomForest	RandomForest_Smote	RandomForest	RandomForest	CatBoost

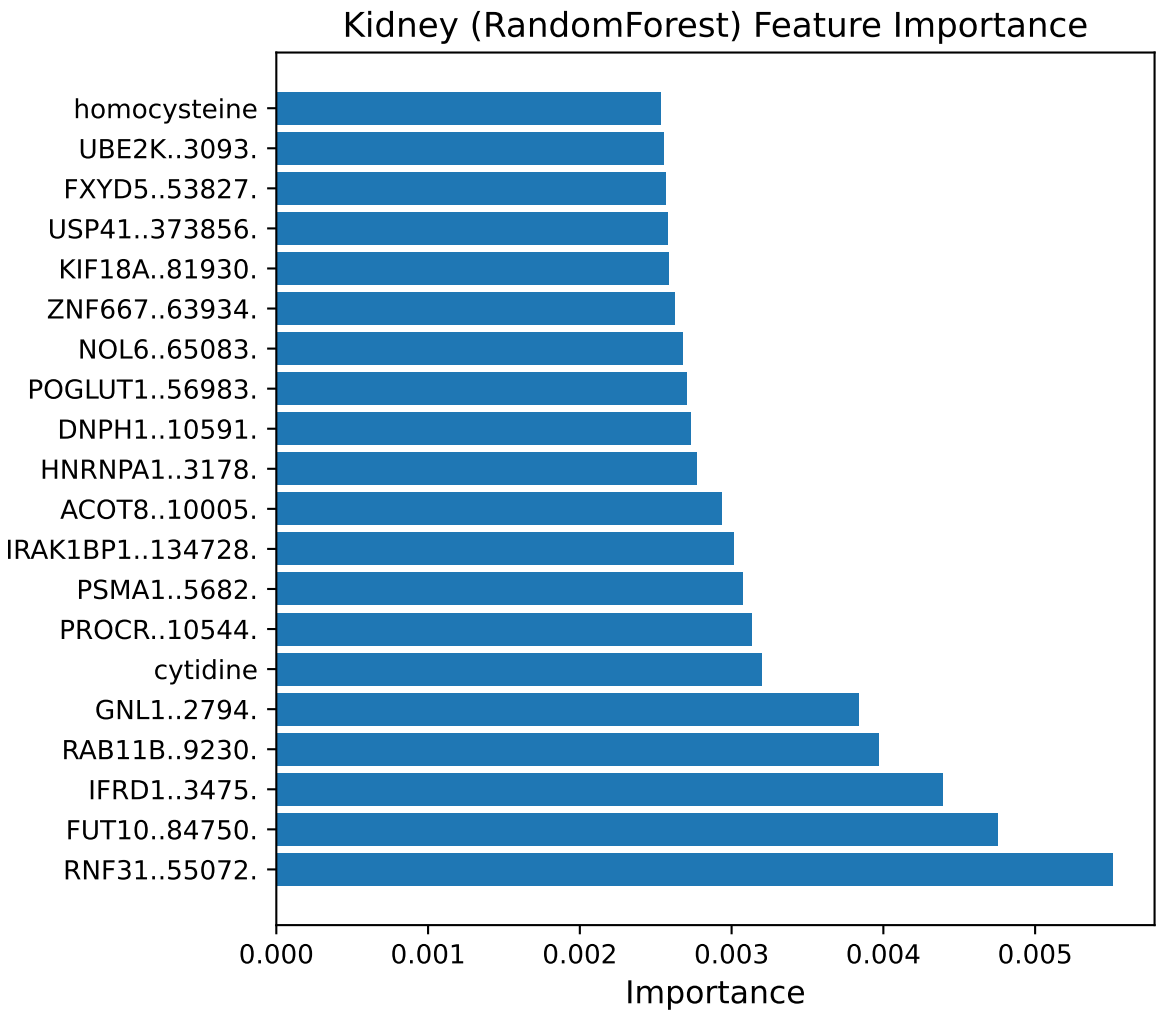
以下為針對不同目標變數預測轉移的最佳模型選擇:

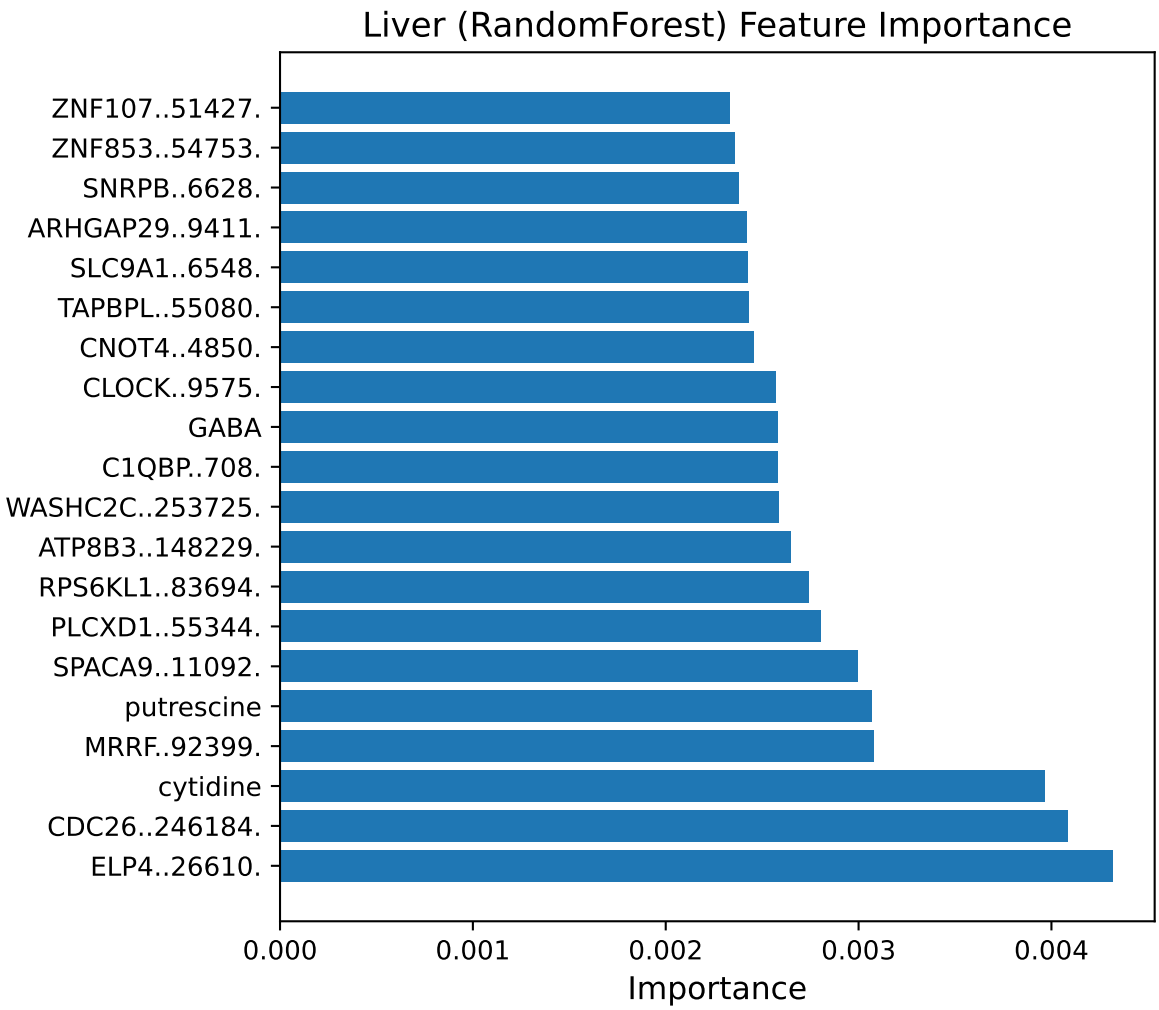
- 對於是否轉移到「骨頭」，隨機森林模型表現最佳，AUC 可達 66.56%。
- 對於是否轉移到「大腦」，結合 SMOTE 與隨機森林模型能夠達到最佳效果，AUC 可達 76.95%。
- 對於是否轉移到「腎臟」，隨機森林模型表現最佳，AUC 可達 71.15%。
- 對於是否轉移到「肝臟」，「隨機森林模型表現最佳，AUC 可達 65.38%。
- 對於是否轉移到「肺臟」，CatBoost 模型表現最佳，AUC 為 56.41%。

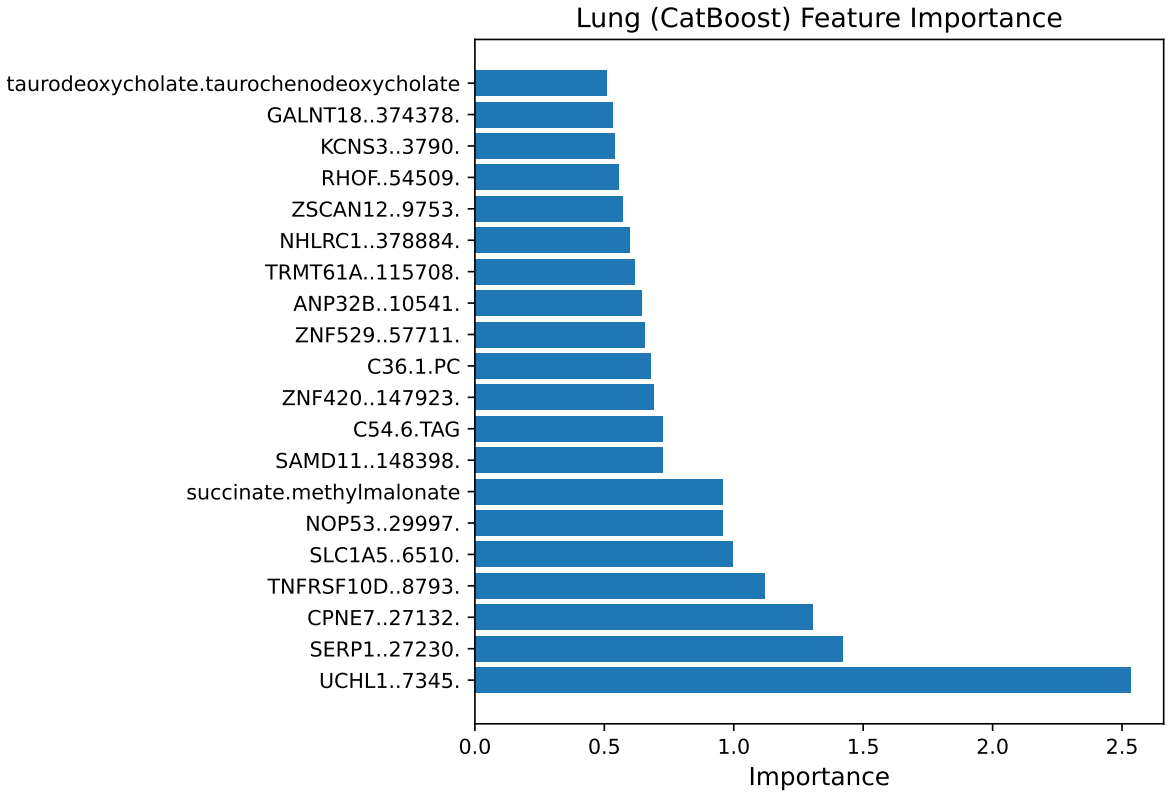
對是否轉移影響最大的前 **20** 個變數




Brain (Smote+RandomForest) Feature Importance
RMDN3..55177.
C54.6.TAG
PRDM5..11107.
C16orf74..404550.
ZNF503..84858.
PBX3..5090.
C58.8.TAG
SNTB1..6641.
CITED2..10370.
ZCWPW1..55063.
C3orf62..375341.
CAPN15..6650.
TRPC1..7220.
asparagine
C56.6.TAG
C54.4.TAG
TCEAL3..85012.
VMP1..81671.
ARHGAP8..23779.
WDR73..84942.
0.000
0.001
0.002
0.003
0.004
0.005
Importance


Kidney (RandomForest) Feature Importance
homocysteine
UBE2K..3093.
FXYD5..53827.
USP41..373856.
KIF18A..81930.
ZNF667..63934.
NOL6..65083.
POGLUT1..56983.
DNPH1..10591.
HNRNPA1..3178.
ACOT8..10005.
IRAK1BP1..134728.
PSMA1..5682.
PROCR..10544.
cytidine
GNL1..2794.
RAB11B..9230.
IFRD1..3475.
FUT10..84750.
RNF31..55072.
0.000
0.001
0.002
0.003
0.004
0.005
Importance


Liver (RandomForest) Feature Importance
ZNF107..51427.
ZNF853..54753.
SNRPB..6628.
ARHGAP29..9411.
SLC9A1..6548.
TAPBPL..55080.
CNOT4..4850.
CLOCK..9575.
GABA
C1QBP..708.
WASHC2C..253725.
ATP8B3..148229.
RPS6KL1..83694.
PLCXD1..55344.
SPACA9..11092.
putrescine
MRRF..92399.
cytidine
CDC26..246184.
ELP4..26610.
0.000
0.001
0.002
0.003
0.004
Importance



- 預測是否轉移至「骨頭」，重要的代謝體為: C34.4.PC
- 預測是否轉移至「大腦」，重要的代謝體為: C54.6.TAG, C58.8.TAG, asparagine, C56.6.TAG, C54.4.TAG
- 預測是否轉移至「腎臟」，重要的代謝體為: homocysteine, cytidine
- 預測是否轉移至「肝臟」，重要的代謝體為: GABA, putrescine, cytidine
- 預測是否轉移至「肺臟」，重要的代謝體為: taurodeoxycholate.taurochenodeoxycholate, C36.1.PC, C54.6.TAG, succinate.methylmalonate