腦瘤影像分割（**Advancing Brain Tumor Segmentation: Exploring Swin UNETR and Transfer Learning on the BraTS 2018 Dataset**）

第2組 組員：林亮瑜、劉冠廷、張維蓁

1. 緒論（**Introduction**）

Brain tumors是一種致命的疾病，分成原發性腦瘤(primary tumors)和繼發性腦瘤(secondary tumors)，原發性腦瘤是指起源於腦部的腫瘤，而繼發性腦瘤為其他癌細胞經過轉移、擴散後形成的腫瘤，其中神經膠質瘤(glioma)是最常見的brain tumor類型，神經膠質瘤是brain glial cells所產生，分為high-grade glioma (HGG)與low-grade glioma (LGG)，high-grade glioma (HGG)是未分化的膠質瘤，腫瘤通常屬於惡性，且患者的預後較差，low-grade glioma (LGG)是分化良好的膠質瘤，惡性腫瘤的傾向低，患者的預後較好 [1]。

腦部核磁共振造影(Magnetic Resonance Imaging, MRI)腫瘤影像分割是醫學影像處理領域中的一個關鍵任務，它在診斷和治療腦瘤方面具有重要的應用價值。隨著醫學影像技術的進步，MRI成為了一種非侵入性高解析度的影像模態，可提供關於腦部結構和組織的詳細資訊。由於腫瘤在不同的影像中的表現不同，通常需要多種影像判斷腫瘤的位置、大小…等，例如：T1、T2、Fluid-attenuated Inversion Recovery (FLAIR)、加顯影劑的 T1 (T1ce)，以強調不同的組織特性和腫瘤擴散區域 [2,3]。然而，由於腦部解剖的複雜性和腫瘤的多樣性，人工分割腦瘤是一項耗時且主觀性強的工作，並且容易受到操作者間的差異性和疲勞等因素的影響。

為了克服這些問題，自動化腦瘤影像分割成為一個熱門的研究領域。它結合了醫學影像學、圖像處理和機器學習等技術，旨在開發出能夠準確且快速地定位和分割腦瘤的算法和方法。自動化分割方法不僅可以提高腦瘤診斷的準確性，還能夠幫助醫生制定更有效的治療計劃、監測腫瘤的生長和治療反應。

本次報告主要目的是想測試BraTS 2021 competition中獲勝的方法 — Swin UNETR [4,5,6]，評估此模型是否同樣也能適用在其他腦瘤資料集上，我們選用BraTS 2018 Dataset進行操作，藉以探討這個模型的外推性，之後進一步以finetune pretrained model [7]優化模型，觀察結果是否比BraTS 2018獲勝者的成效更好 [8]，期望以transfer learning的方式有效地改進模型的預測效果。

2. 研究材料與方法（**Methods**）

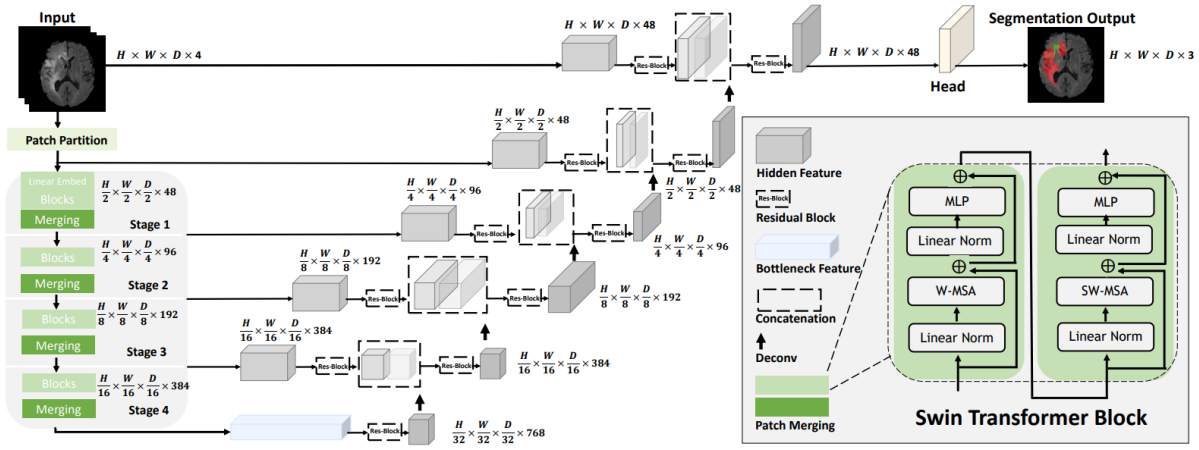
2.1. Swin UNETR[9]

2.1.1 Model Architecture

Swin UNETR架構如Fig. 1所示，主要是由encoder、bottleneck、decoder、skip connection四個部分所組成，模型是以U-Net作為基礎架構，主要差異是在encoder的部分多加了Swin Transformer Block，而Swin Transformer Block會以不同分辨率提取特徵，透過skip connection在每個分辨率上連接到decoder，最後得到segmentation

output。

Fig. 1. Swin UNETR的架構概述



2.1.2 Data Preprocessing and Augmentation

我們將非零voxels的input影像進行normalization，使其具有zero mean與unit standard deviation，Fig. 2, 3中顯示出intensity從原本範圍(-175.0, 250.0)變成(0.0, 1.0)，之後也將3個軸以0.5的機率隨機進行鏡向翻轉，如Fig. 3所示。另外，在data augmentation的部分，我們會先隨機將每個image channel的intensity scaling至範圍(0.9, 1.1)，scaling公式為v = v * (1 + factor)，Fig. 3顯示factor設定在0.1，之後再進行隨機shifting至範圍(-0.1, 0.1)中，shifting的範圍是在(-offsets, offsets)之間，我們將offsets設定在0.1，如Fig. 3所示 [10]。

Fig. 2. Data Normalization：a_min denotes intensity original range min, a_max denotes intensity original range max, b_min denotes intensity target range min, b_max denotes intensity target range max

```
parser.add_argument("--a_min", default=-175.0, type=float, help="a_min in ScaleIntensityRanged")
parser.add_argument("--a_max", default=250.0, type=float, help="a_max in ScaleIntensityRanged")
parser.add_argument("--b_min", default=0.0, type=float, help="b_min in ScaleIntensityRanged")
parser.add_argument("--b_max", default=1.0, type=float, help="b_max in ScaleIntensityRanged")
```

Fig. 3. Random axis mirror flip and Intensity adjustment

```
transforms.RandFlipd(keys=["image", "label"], prob=0.5, spatial_axis=0),
transforms.RandFlipd(keys=["image", "label"], prob=0.5, spatial_axis=1),
transforms.RandFlipd(keys=["image", "label"], prob=0.5, spatial_axis=2),
transforms.NormalizeIntensityd(keys="image", nonzero=True, channel_wise=True),
transforms.RandScaleIntensityd(keys="image", factors=0.1, prob=1.0),
transforms.RandShiftIntensityd(keys="image", offsets=0.1, prob=1.0),
transforms.ToTensord(keys=["image", "label"]),
```

2.2. 資料集（Datasets）

2.2.1 BraTS 2018與BraTS 2021 Dataset的比較

由於Swin UNETR在資料前處理方面進行了大量工作，我們預期它具有良好的外推能力，在研究模型外推性之前，要先確定兩者的資料集的差異，因此我們選用了BraTS 2018 Dataset進行操作。從Kaggle上取得BraTS 2018與BraTS 2021 Dataset [11,12]，BraTS challenge主要是透過所提供的3D MRI資料集和醫生注釋的ground truth tumor segmentation labels訓練model辨別出tumor的位置。BraTS 2018 training dataset中包含285 cases，分成210個HGG與75個LGG，另外也有提供survival data用以預測存活率。而BraTS 2021 training dataset中包含1251 cases，並未區分HGG、LGG。BraTS 2018與BraTS 2021 Dataset的差異如Table 1所示。

兩個資料集的影像大小皆為240 × 240 × 155，每個case都有4種3D MRI modalities (T1, T1ce, T2 and FLAIR)，而ground truth annotation包含3個tumor sub-regions: enhancing tumor、peritumoral edema、necrotic and non-enhancing tumor core，這些annotation會被組合成3個nested sub-regions: Whole tumor (WT)、Tumor core (TC)、Enhancing tumor (ET)。

Table 1. BraTS 2018與BraTS 2021 Dataset的差異

	BraTS 2018	BraTS 2021
是否有區分HGG/LGG？	O	X
Training Data 數	Total : 285 HGG(210)、LGG(75)	Total : 1251
Validation Data 數	66	219
Survival Data	O	X

2.2.2 影像呈現（**Dataset Image**）

在這部分，我們將呈現先前提到的兩個資料集的內容並利用視覺化技巧展示它們之間的差異。首先，我們會從2018年和2021年的資料集中選取一個樣本，以呈現MRI的影像切片的影像內容，下圖是將4種3D MRI造影方式(T1, T1ce, T2 and FLAIR)分別作圖，而Seg為醫生所標註的ground truth labels。Fig. 4為2018年資料集中的Brats18_2013_2_1，屬於HGG分類。Fig. 5為2018年資料集中的Brats18_2013_1_1，屬於LGG分類。Fig. 6為2021年資料集中的BraTS2021_00495。可以看到2021年的資料品質較高，畫質較清晰，而2021年的資料品質較差，畫質較模糊。

Fig. 4. Brats18_2013_2_1 (2018 Dataset)

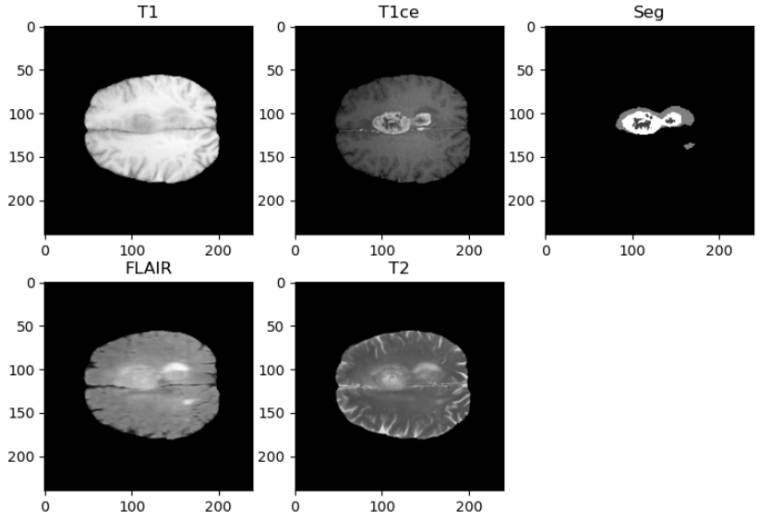


Fig. 5. Brats18_2013_1_1 (2018 Dataset)

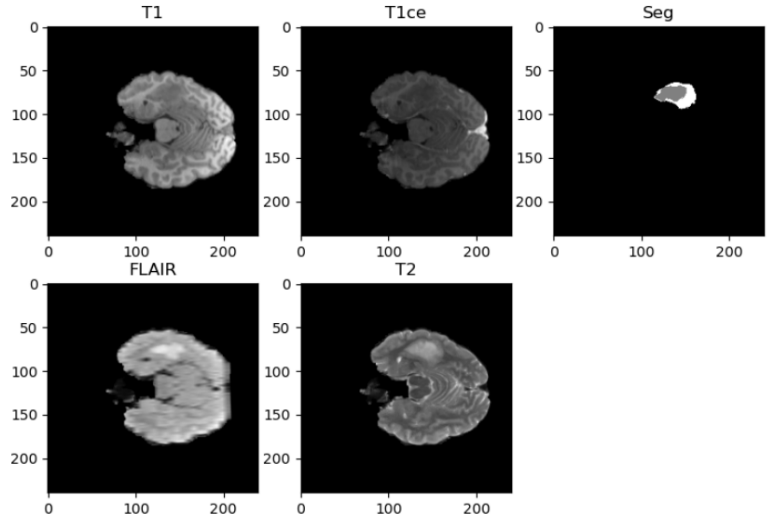
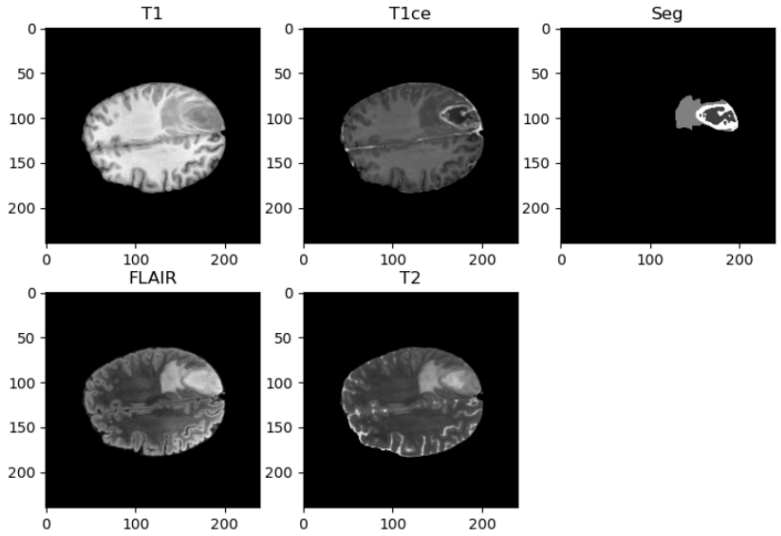


Fig. 6. BraTS2021_00495 (2021 Dataset)

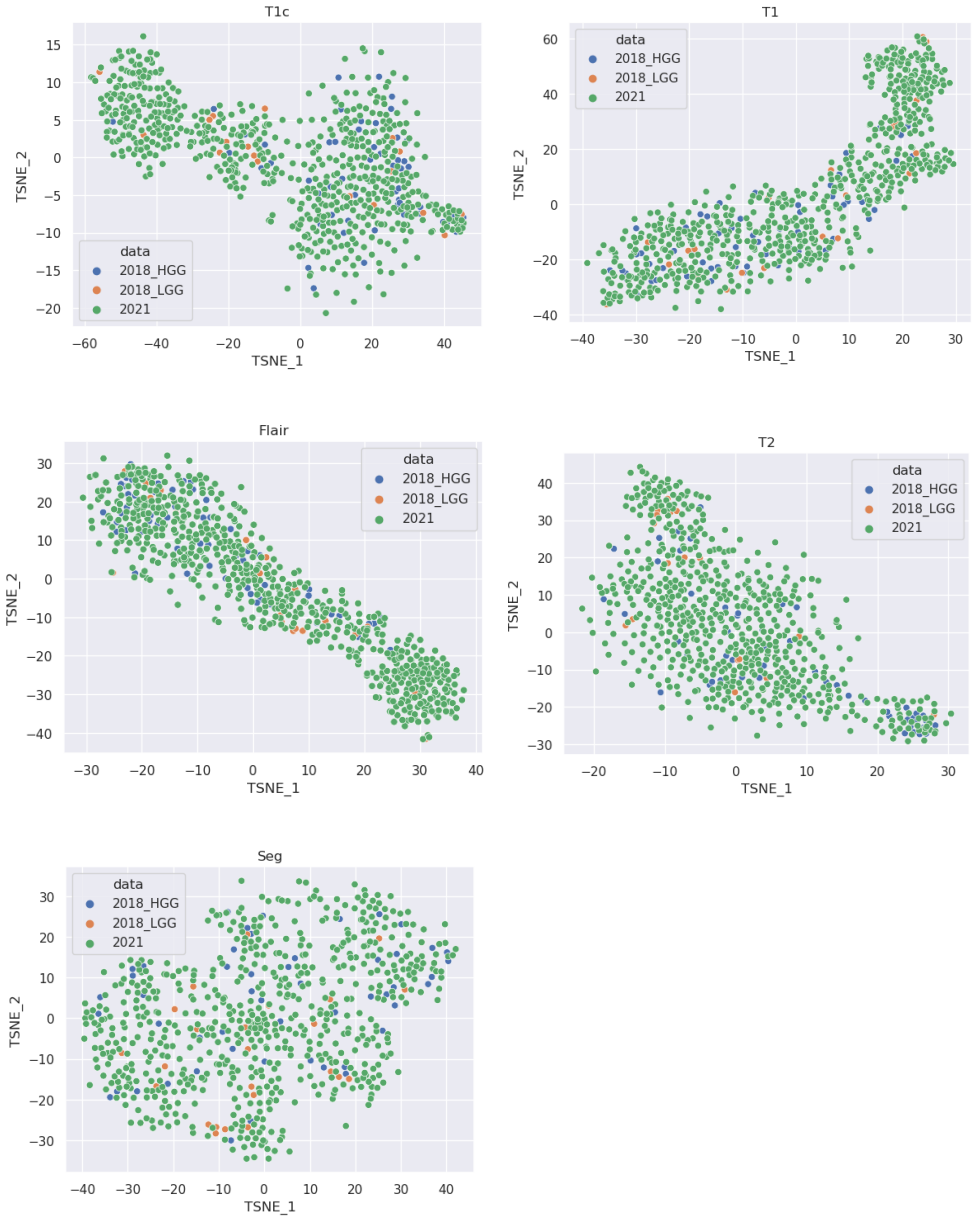


2.2.3 資料視覺化（**Data visualization**）

接下來，我們將運用t-SNE（t-distributed stochastic neighbor embedding）方法[13]，把高維度的3D影像資料轉換為更易於作圖理解的二維表示形式。通過觀察它們在二維空間中的分布和相對位置，評估這兩個資料集之間的異同，這樣可以幫助我們理解這兩個資料集之間的關係、群集結構以及可能存在的模式或趨勢。

受限於記憶體的關係，本階段使用一半的資料降維作圖，分別將Brats18 (HGG)、Brats18 (LGG)、Brats2021以隨機方式篩選一半的資料，*n_components*設定為2、*random_state*設定為42進行t-SNE降維處理。

Fig. 7. 兩個資料集影像特徵在經t-SNE降維後皆重疊



3. 模型測試結果（**Results**）

3.1. Swin UNETR預訓練模型在**BraTS2018**的表現

首先本次實驗直接應用Swin UNETR中表現最好的預訓練模型(Fold1)不經訓練直接測試BraTS2018的資料，程式碼參見下圖。結果如下表所示，可以看到儘管模型訓練過程中使用許多不同的資料前處理方式，2021的新模型直接使用在2018資料集上的表現明顯不如原測試資料，平均準確性為0.31，本次實驗也將之測試於2021資料，作為對照再現了該文章所提出的準確度0.90，排除掉是程式碼錯誤的可能性。由此得知預訓練模型無法準確預測2018年的數據集。

Fig. 8. Code example of pre-trained model testing

```
(swin) 109304021@c007:~/swin$ python main.py --resume_ckpt --workers 0 --val_every 1 --fold 3 --json_list=jsons/brats2018_fold.json --data_dir=2018_trai
ning/MICCAI_BraTS_2018_Data_Training/ --distributed --max_epochs 1
train: 231
validation: 54
0  gpu 0
Batch size is: 1 epochs 1
Using pretrained weights
Total parameters count 62191941
Writing Tensorboard logs to  ./runs/test
0 Mon May 29 14:24:03 2023 Epoch: 0
/home/109304021/anaconda3/envs/swin/lib/python3.7/site-packages/monai/metrics/meandice.py:78: UserWarning: y_pred should be a binarized tensor.
  warnings.warn("y_pred should be a binarized tensor.")
Val 0/1 0/54 , Dice_TC: 0.42521814 , Dice_WT: 0.9305793 , Dice_ET: 0.3455174 , time 20.14s
Val 0/1 1/54 , Dice_TC: 0.529399 , Dice_WT: 0.6273658 , Dice_ET: 0.51985264 , time 13.02s
Val 0/1 2/54 , Dice_TC: 0.35293266 , Dice_WT: 0.45141873 , Dice_ET: 0.34656844 , time 12.80s
Val 0/1 3/54 , Dice_TC: 0.2646995 , Dice_WT: 0.38517028 , Dice_ET: 0.25992632 , time 12.74s
Val 0/1 4/54 , Dice_TC: 0.3357696 , Dice_WT: 0.41443592 , Dice_ET: 0.3586741 , time 12.79s
Val 0/1 5/54 , Dice_TC: 0.31555295 , Dice_WT: 0.37798536 , Dice_ET: 0.32788244 , time 12.88s
```

Table 2. Testing results of pre-trained model

	Dice_TC	Dice_WT	Dice_ET	Accuracy
Fold-0 (65)	0.28285	0.53255	0.22598	0.34713
Fold-1 (64)	0.22941	0.32770	0.17163	0.24291
Fold-2 (47)	0.29992	0.42889	0.23711	0.32198
Fold-3 (54)	0.29015	0.47305	0.24850	0.33724
Fold-4 (55)	0.29225	0.40019	0.21032	0.30092
Average	**0.27892**	**0.43248**	**0.21871**	**0.31004**
Fold-1 (2021 best)	0.90075	0.92517	0.88792	0.90461

3.2. Finetuned model performance

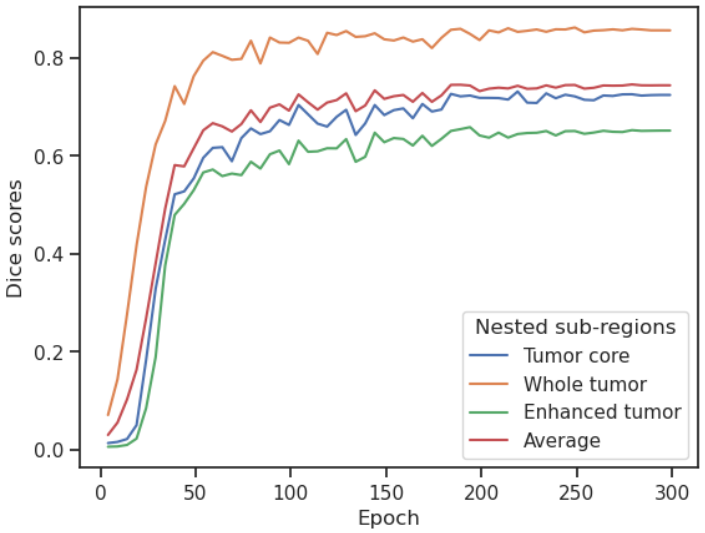
接下來，我們使用BraTs 2018 Training dataset 對Swin UNETR進行微調。首先，我們先將Training data切分成5-fold，其中取 fold-0, 2, 3, 4 (n = 221) 當作訓練資料，而 fold-1 (n = 64)作為驗證資料，並利用Swin UNETR中表現最好的預訓練模型 (Fold1)進行微調 (resize:128x128x128, batch size: 1, epochs: 300)，微調模型程式碼如下圖 (**Fig. 9**)。結果如下圖 (**Fig. 10-a, b**) 所示，Swin UNETR模型的Accuracy從原本的0.31大幅進步到0.746，其中各種subregion的Dice score為Dice_TC: 0.726 , Dice_WT: 0.860 , Dice_ET: 0.653。

Fig. 9 Code example of model finetuning

```
(swin) 109304021@c007:~/swin$ python main.py --json_list=jsons/brats2018_fold.json --val_every=5 --data_dir=2018_training/MICCAI_BraTS_2018_Data_Trainin
g/ --noamp --pretrained_model_name=model.pt --pretrained_dir=pretrained_models/fold1_f48_ep300_4gpu_dice0_9059 --fold=1 --roi_x=128 --roi_y=128 --roi_z=
128  --in_channels=4 --spatial_dims=3 --use_checkpoint --feature_size=48 --sw_batch_size=4 --workers=1 --distributed
Found total gpus 1
train: 221
validation: 64
0  gpu 0
Batch size is: 1 epochs 300
Total parameters count 62191941
Writing Tensorboard logs to  ./runs/test
0 Mon May 22 10:46:01 2023 Epoch: 0
Epoch 0/300 0/221 loss: 0.9077 time 27.78s
Epoch 0/300 1/221 loss: 0.9526 time 3.19s
Epoch 0/300 2/221 loss: 0.9524 time 3.17s
```

Fig. 10-a Loss and average dice score of three sub-regions of finetuning

Fig. 10-b Dice score of three sub-regions of finetuning



4. 討論（Discussion）

根據先前競賽冠軍的模型表現可以發現無論是在2018資料或是2021資料最佳的模型表現都可以達到平均Dice score 0.8以上甚至達到0.9以上的成果 [9,14]。然而卻觀察到pre-trained model在2018資料上表現不如預期，沒有外推性，準確性僅有0.310，主要推測造成預訓練模型無法準確判斷先前資料集的主因是影像品質所導致的。可能由於2018年的影像資料品質較差，包含噪點、模糊或低對比度等問題，這使得模型難以從中獲得準確的特徵，儘管影像經T-SNE降維後的特徵空間上重疊但可能是降維過程丟失了過多資料，這兩組資料影像本質仍不相同。除此之外，也可能是因為不同醫生之間的解讀存在差異，導致兩個數據集之間存在不同的標註標準。這種差異性阻礙了模型從2018年醫生的判斷中學習，使其難以捕捉到相應的特徵和模式。因此，這些因素共同作用，限制了預訓練模型對於2018年資料集的準確性。為了克服這些問題，可以通過對預訓練模型進行微調，以提高其對於先前資料集的準確性和泛化能力，本次實驗之微調後準確性達到0.746。

5. 結論（Conclusion）

質性分析和T-SNE降維方法在2018年和2021年的數據集上呈現相似的特徵，但從影像來看可以看出影像品質的差異。儘管這兩個資料集很類似，但預訓練模型無法準確預測2018年的數據集。通過對預訓練模型進行微調，可以將準確性從0.310提升至0.746（2018年最佳結果為0.859）。主要推測造成預訓練模型無法準確判斷先前資料集的主因是影像品質所導致的，另外也可能是因為不同醫生之間的解讀存在差異，以致於兩個數據集之間存在不同的標註標準。透過本次實驗可以看到儘管資料極度相似，新的模型表現的很好但應用於舊有資料時遷移學習仍是必要的過程。

References

1. Zacharaki, E.I., Wang, S., Chawla, S., Soo Yoo, D., Wolf, R., Melhem, E.R., Davatzikos, C.: Classification of brain tumor type and grade using mri texture and shape in a machine learning scheme. Magnetic Resonance in Medicine: An Official Journal of the International Society for Magnetic Resonance in Medicine 62(6), 1609–1618 (2009)
2. Grover, V.P., Tognarelli, J.M., Crossey, M.M., Cox, I.J., Taylor-Robinson, S.D., McPhail, M.J.: Magnetic resonance imaging: principles and techniques: lessons for clinicians. Journal of clinical and experimental hepatology 5(3), 246–255 (2015)
3. https://medebm.blogspot.com/2018/01/magnetic-resonance-imaging-mri.html
4. https://rdcu.be/dc2vy
5. https://github.com/Project-MONAI/research-contributions/tree/main/SwinUNETR/BRATS21
6. https://colab.research.google.com/github/Project-MONAI/tutorials/blob/main/3d_segmentation/swin_unetr_brats21_segmentation_3d.ipynb#scrollTo=953VEvWgmDA8
7. https://jason-chen-1992.weebly.com/home/fine-tuning
8. https://rdcu.be/dc2uH
9. Hatamizadeh, Ali, et al. "Swin unetr: Swin transformers for semantic segmentation of brain tumors in mri images." Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries: 7th International Workshop, BrainLes 2021, Held in Conjunction with MICCAI 2021, Virtual Event, September 27, 2021, Revised Selected Papers, Part I. Cham: Springer International Publishing, 2022.
10. https://docs.monai.io/en/stable/transforms.html#randshiftintensity
11. https://www.kaggle.com/datasets/sanglequang/brats2018
12. https://www.kaggle.com/datasets/dschettler8845/brats-2021-task1
13. Van der Maaten, Laurens, and Geoffrey Hinton. "Visualizing data using t-SNE." Journal of machine learning research 9.11 (2008).
14. Myronenko, Andriy. "3D MRI brain tumor segmentation using autoencoder regularization." Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries: 4th International Workshop, BrainLes 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 16, 2018, Revised Selected Papers, Part II 4. Springer International Publishing, 2019.