공학석사 학위논문

# 다중 테스크 학습을 통한 DWI-FLAIR 영상의 신호 불일치를 이용한 뇌졸중 onset 예측에 관한 연구: 외부 데이터셋들을 이용한 모델 검증

Identifying stroke onset time on DWI-FLAIR mismatch using deep learning with multi-task learning: validation with external datasets

울산대학교대학원

의 과 학 과

남 유 진

# 다중 테스크 학습을 통한 DWI-FLAIR 영상의 신호 불일치를 이용한 뇌졸중 onset 예측에 관한 연구: 외부 데이터셋들을 이용한 모델 검증

지도교수 김 남 국

이 논문을 공학석사 학위 논문으로 제출함

2023 년 08 월

울산대학교 대학원

의 과 학 과

남 유 진

남유진의 공학석사학위 논문을 인준함

심사위원 김 남 국 (인)

심사위원 이 준 구 (인)

심사위원 이 은 재 (인)

울 산 대 학 교 대 학 원

2023 년 08 월

**Abstract**

The treatment of acute ischemic stroke, which has a strong correlation between the time clock and tissue progression, heavily relies on the onset time. However, determining the exact timing of onset can be challenging due to factors such as stroke occurrence during sleep or unclear onset. To address this issue, previous studies have utilized the signal mismatch between DWI and FLAIR images to estimate the elapsed time after stroke onset.

In this study, our goal is to develop a technique for determining the time elapsed after stroke using deep learning with multi-task learning (MTL) approach based on MRI image biomarkers. We aim to evaluate the performance of this technique using external validation datasets.

The proposed model takes a combination of ADC, FLAIR, and DWI-b1000 volumes as input, and performs simultaneous prediction of stroke onset time and infarct segmentation for time windows of 3 hours, 4.5 hours, and 6 hours using MTL. The backbone of the model is the 3D patch based SwinUNETR model, with the addition of auxiliary classifiers in the bottleneck layer. A voting ensemble of patch-level classifiers is performed at the patient level for classification. The auxiliary classifiers count and vote for onset identification based on the presence of infarction in patches using the predicted mask.

For the classification task, we compared the performance of our model with 3D DenseNet and 3D patched DenseNet using internal validation, using AUC, accuracy, specificity, and sensitivity as evaluation metrics. We also compared the performance of onset classification for 4.5 hours using two external validations. For the segmentation task, we compared the performance of DiNTS, UNETR, nnUNET, and our proposed model using IOU and Dice coefficient as evaluation metrics. Overall, our proposed model showed superior performance in classification compared to the

compared models, and it also demonstrated similar or improved segmentation results compared to nnUNET. Additionally, when predicting onset time, the model showed attention to the lesion by extracting Grad-CAM from the encoder, indicating its focus on the infarction during training. We provided both quantitative and qualitative evaluations across multiple segmentation and classification tasks. This MTL-based model showed better performance in identifying stroke onset time in internal and two external validations, demonstrating its potential for potential clinical use. The proposed model has the potential to assist in performing thrombolysis therapy at the appropriate time for stroke patients with unclear onset time by predicting the onset time accurately.

## Abbreviations

AIS (acute ischemic stroke)

IV-tPA (intravenous recombinant tissue plasminogen activator)

DWI (diffuse-weighted imaging)

ADC (apparent diffusion coefficient)

FLAIR (fluid- attenuated inversion recovery)

MTL (multi-task learning)

RPN (region proposal network)

TSS (stroke symptom onset)

Swin UNETR (Swin U-Net Transformer)

SPM (Statistical Parametric Mapping)

PPV (positive predictive value)

NPV (negative predictive value)

ROC (receiver operating characteristic)

AUCs (areas under the ROC curves)

DSC (dice similarity coefficient)

IOU (intersection over union)

Grad-CAM (gradient-weighted class activation map)

## Contents

## Contents of Tables

**Contents of Figures**

## Introduction

Stroke is the second leading cause of death and disability worldwide, with more than 13 million cases reported each year [1]. One in four people experience stroke throughout their lifetime and ischemic strokes comprise 62% of all strokes, which often result in irrecoverable neurological sequelae, morbidity, and mortality [2]. For acute ischemic stroke (AIS) patients, vascular recanalization is the most important factor in deciding the patients' prognosis, [3] and although the role of mechanical thrombectomy has expanded, proper administration of intravenous recombinant tissue plasminogen activator (IV-tPA) thrombolysis within 4.5 hours after symptom onset is critical for improving patients' functional outcomes [4-7]. To choose right candidates for intravenous thrombolysis, identifying the exact time of symptom onset is a very important issue, but about 25% of AIS event are known to occur during sleep, as we call it wake-up stroke, and thus it is difficult to determine the exact symptom onset time [8-11]. To address this problem, many studies have confirmed that the symptom onset time can be predicted using the inconsistency between diffuse-weighted imaging (DWI) and fluid- attenuated inversion recovery (FLAIR) imaging (visible acute lesion on DWI and lack of prominent hyperintensity on FLAIR in the corresponding region) [12-15]. For instance, the water diffusion change due to brain tissue ischemia can be detected within minutes from the symptom onset through reduction of apparent diffusion coefficient (ADC) values [16-18], while the increase in water diffusion can be visualized in T2 images as high intensity after 1-4 hours from the onset [18, 19]. Accordingly, mismatch of multiparametric MR imaging has been used to estimate the elapsed time of 4.5 hours after the onset of the disease in ischemic stroke patients and to determine the stroke age in AIS patients whose onset time is unknown [20-22]. In practice, the 2019 update on the ASA/AHA guideline for the early management of patients with acute stroke newly recommended that DWI-FLAIR mismatch can be used for choosing patients who may benefit from IV alteplase in the setting of unclear time of symptom onset based on

the WAKE-UP trial [6]. However, human readings of mismatch DWI-FLAIR signals generally tend to have low-to-moderate intra- and inter-rater variability, which brings large differences in specificity and sensitivity for predicting whether the symptom onset time is within 4.5 hours or not [20, 21, 23-25].

*Single task learning with identifying AIS*

There have been numerous studies using machine learning and deep learning methods to identify stroke onset. In terms of machine learning, one approach involved extracting 89 vector features from each image sequence and using them to develop three models: logistic regression, support vector machine, and random forest [26]. Another study utilized a two-modal MR image (DWI, FLAIR) along with a machine learning classifier trained on lesion segmentation from DWI to determine time since stroke [27]. These models shared the commonality of extracting features based on lesion information and lacking external validation, which may raise concerns regarding their robustness and reproducibility. Several deep learning methods have also been proposed. In this particular paper, the focus was on developing machine learning models to classify stroke patients with unknown time since stroke, and a deep learning model was introduced to extract hidden representations from perfusion-weighted MRI images, resulting in improved classification. However, overall performance in previous studies has been suboptimal, and similar to the machine learning approaches, external validation was lacking [28]. Another study presented a deep learning model for automatically determining time since stroke, but it exhibited discrepancies in performance compared to external validation [29].

*Multi-task learning*

Multitask learning (MTL) is a learning paradigm that aims to improve the generalization performance of multiple related tasks by leveraging their relational information [30]. While

MTL has not been used for predicting the onset of AIS so far, it has been applied to other medical images. In the case of 2D ultrasound images, a modified U-Net-based multi-feature guided CNN architecture was proposed. They added a classification branch to the bottom of the U-Net for simultaneous segmentation and classification of bone surfaces in 2D ultrasound data [31]. Additionally, a joint training network combining a region proposal network (RPN) backbone with three sub-networks for detection, classification, and segmentation of skin lesions was proposed [32]. However, when processing CT, MRI, and other medical images, using 2D CNNs with single CT slices as input fails to leverage the temporal context from adjacent slices. On the other hand, 3D convolution kernels can utilize the inter-slice context from volumetric inputs. 3D CNNs can leverage both spatial and temporal features compared to 2D CNNs, resulting in improved performance. Therefore, our model proposes a 3D patch-based transformer model using multi-task learning.

*Transformer with vision tasks*

Previous studies on AIS onset time prediction have focused on extracting infarct features, which often involves a complex and cumbersome process. To overcome this challenge, deep learning models can be employed to automatically segment infarcts and extract relevant features. Ensemble models based on simple U-Net architectures have demonstrated promising results in brain infarct segmentation. For example, Kamnitsas et al. [33] proposed a robust segmentation model by combining outputs from various CNN-based models, such as 3D U-Net, 3D FCN, and Deep Medic. Zhou et al. [34] introduced an ensemble of CNN-based networks that leveraged multi-scale contextual information through an attention block. Isensee et al. [35] developed the nnU-Net model, showcasing competitive performance in this field.

Transformer-based models, which have gained attention in natural language processing and computer vision, have also made an impact in medical image analysis [36-

38]. Vision Transformers (ViTs) [38] have particularly excelled in computer vision benchmarks, thanks to their self-attention module that captures long-range dependencies between token embeddings and enables effective contextual representations [39]. UNETR [40] stands out as the first methodology to utilize a ViT as an encoder without relying on a CNN-based feature extractor. Although UNETR has demonstrated excellent accuracy and efficiency in various medical image segmentation tasks, transformer-based models employing Swin Transformers have been introduced [41, 42]. Swin Transformers are hierarchical vision transformers that compute self-attention using a shifted window partitioning scheme, making them suitable for tasks that require multi-scale feature extraction. Therefore, we propose a model that incorporates an auxiliary classifier into the Swin UNet Transformer (Swin UNETR) [43]. This model performs infarct segmentation and simultaneously predicts AIS onset time based on the segmented infarcts.

*Objectives*

1. Automatic end-to-end algorithm without human effort using segmentation and classification method with multi-task learning.
2. Demonstrate that our method outperforms human reading method, the previous radiomics-based ML, and DL method.
3. We visually demonstrated the explainability of our method by using activation maps, focusing on the infarction of the DWI-FLAIR mismatch.
4. Our model was validated for robustness and reproducibility through two external validations.

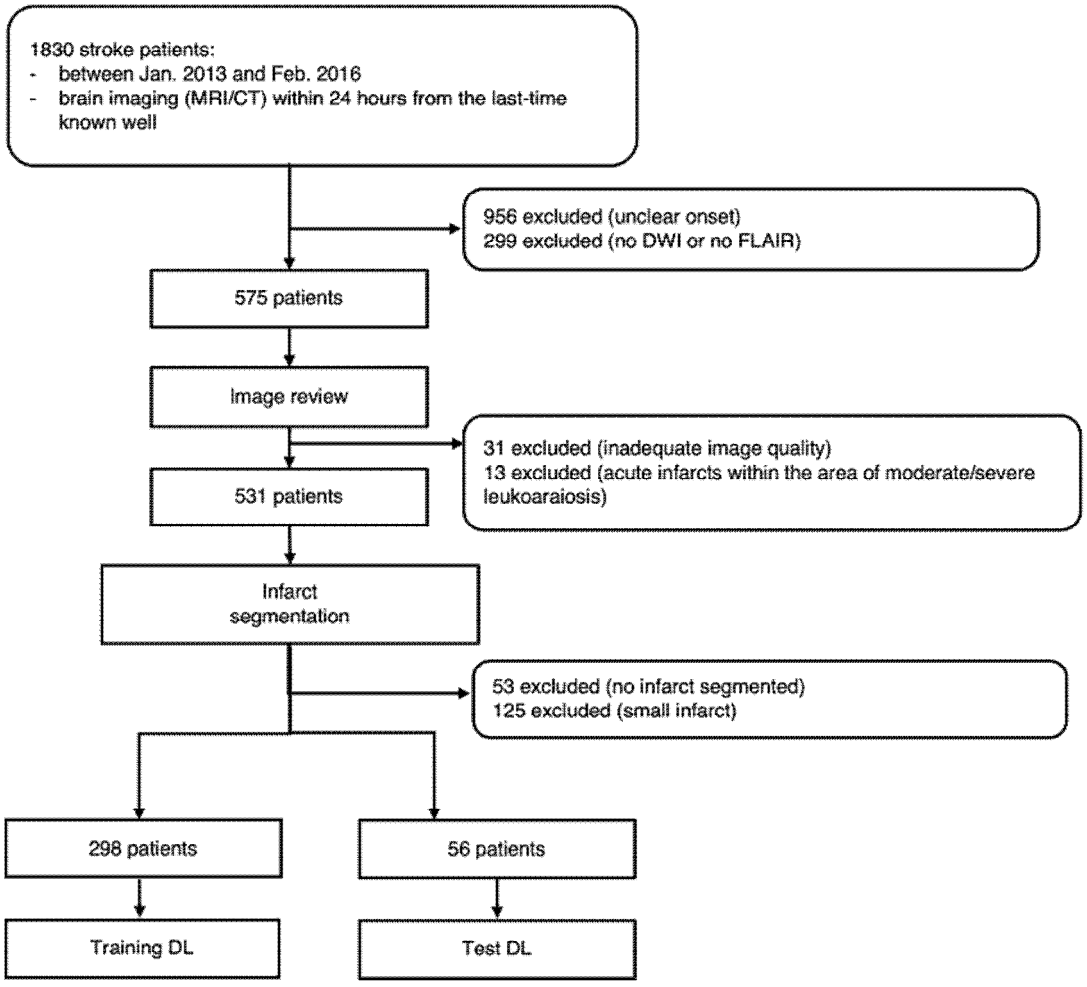
**Onset Time Prediction with DWI-FLAIR Mismatch using Multi-Task Learning**

## 1. Dataset

### 1.1. MRI protocols

The study used three 1.5T clinical systemic MRI scanners from different vendors (Avanto by Siemens, Signa by GE Medical Systems, and Archieva by Philips Medical Systems), each equipped with standard head coils. The acute stroke MRI protocol for each patient involved obtaining DWI and FLAIR sequences, with the ADC values being automatically generated from the DWI scans using the scanners' built-in software.

Figure 1. Data flowchart of training and validation dataset



### 1.2. Data description

This study retrospectively evaluated consecutive patients with acute ischemic stroke who visited Asan Medical Center (Seoul, South Korea), between January 2013 and February 2016. The patients were included if their symptom onset time was clearly recorded, and if they underwent an MRI scan that included both DWI and FLAIR sequences within 24 hours of symptom detection. The study aimed to target a sample size of 355 patients based on previous research using deep learning techniques in acute stroke imaging. Out of 1830 patients in the stroke registry, 575 met the criteria for analysis as shown in Figure 1. Among the 575 patients identified for the study, 43 were excluded for various reasons, such as low-quality MRI images, acute lesions within areas of extensive leukoaraiosis, non-segmented lesions, or lesions smaller than 1cc. However, the presence of leukoaraiosis on FLAIR was not a reason for exclusion. Also, the study excluded 53 patients who had lesions that were too small to be segmented, as well as 125 patients with lesions smaller than 1 cc (=103 $mm^3$). This was because image texture features based on statistical measurements cannot be accurately defined in small regions of interest, and noise can affect the precision of the measured textures [44]. However, patients with multiple ischemic lesions were included in the analysis, and the signals from each stroke lesion larger than 1 cc were treated as if they were signals from a single lesion.

Finally, a total of 355 patients were included for the deep learning analysis. The included patients were split into training and validation sets at an 8:2 ratio. To ensure that the distributions of each set were balanced, patients were categorized into eight different time bins ranging from symptom onset times of 0 to 90 minutes up to 1081 to 1440 minutes. Within each time bin, 85% of patients were randomly assigned to the training set, and 15% were assigned to the test set. The study protocol was approved by the institutional review board of the center, which waived the need for written informed consent due to the retrospective nature of the study.

Table 1. Distribution of clinical demographics for the datasets: internal and external validation datasets

| | Internal dataset | | External validation 1 | External validation 2 |
|---|---|---|---|---|
| | Training set (n = 298) | Test set (n = 56) | Test set (n = 104) | Test set (n = 346) |
| Age (years) | 63 (55–73) | 67 (55–71) | 71 (61-80) | 69 (61-81) |
| Female | 86 (28%) | 20 (36%) | 49 (47%) | 88 (52%) |
| Onset to MRI (minutes) | 270 (152–715) | 240 (142–448) | 111 (40-642) | 224 (107-781) |
| Admission NIHSS | 4 (2–10) | 5 (2–12) | 9 (5-16) | 7 (3-17) |
| Within 4.5-hour window (%) | 153 (51%) | 24 (43%) | 64 (61%) | 199 (53%) |

Table 1 presents a comprehensive overview of the clinical information for the datasets used in this study, including the internal training, internal validation, external validation 1, and external validation 2 datasets. The average age distribution is comparable across all datasets, indicating a comparable representation of age groups. Among the datasets, external validation 1 exhibited a higher percentage of females, accounting for 47% of the population, while the other datasets showed a relatively balanced distribution of females ranging from 24% to 36%. The onset to MRI time was consistent across most datasets, except for external validation 1, which demonstrated a higher average time of 111 minutes and an interquartile range spanning from 40 to 642 minutes. In contrast, the other datasets demonstrate average times in the 200s range. The range of onset to MRI time varied, with internal training and external validation 2 displaying a broader range from the 100s to 700s, while the internal test set had a narrower range of 142 to 448 minutes. Notably, the internal training dataset demonstrated a well-balanced representation of cases within the 4.5-hour treatment window, accounting for approximately 50% of the cases. Similarly, the other validation datasets also demonstrate a comparable distribution of cases within this critical treatment window,

highlighting the balanced nature of the datasets utilized in this study.

### 1.3. Human visual assessment of DWI-FLAIR mismatch

Two stroke neurologists, Eun-Jae Lee and HanBin Lee, independently visually assessed DWI-FLAIR mismatches for each patient in the test set. Disagreements were resolved through consensus. The assessment was performed using identical high-resolution screens on two workstations in our center's picture archiving and communicating system (PetaVision; Asan Medical Center, Seoul, South Korea). The observers had the flexibility to adjust window and level settings to optimize contrast for lesion identification. The presence of acute ischemic lesions on FLAIR images was determined by investigators who had knowledge of the lesion presence on DWI. A lesion was considered visible on FLAIR imaging (FLAIR positive) when traceable parenchymal hyperintensity corresponding to the acute ischemic lesion on DWI was present. A DWI-FLAIR mismatch was defined as the presence of a visible acute ischemic lesion on DWI but the absence of traceable parenchymal hyperintensity in the corresponding region on FLAIR imaging (FLAIR negative).

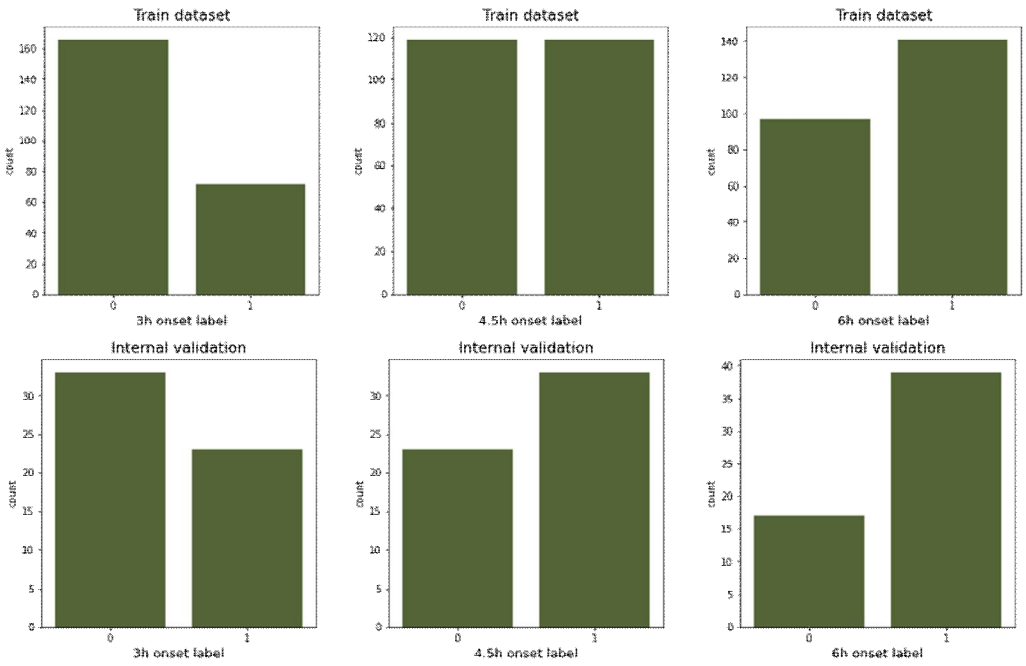
### 1.4. Infarct segmentation

In the study, we analyzed anonymized FLAIR images and ADC maps derived from DWI images. The ADC maps were used to automatically segment infarct regions through normalized absolute thresholding. For each subject's ADC map, a quantile curve of ADC intensities within the brain mask was created. By identifying the intersection point between two tangent lines with maximum and minimum differential coefficients on the quantile-intensity curve, we were able to normalize the ADC maps. This normalization process led to the convergence of intensities in normal white matter tissues to one. Following normalization, each ADC map was thresholded at the optimal absolute value of 0.845.

### 1.5. Two datasets of external validation

To validate the reproducibility and robustness of our model, we used two datasets. The first external validation set consists of data collected from Asan Medical Center from 2017 to 2021, comprising 104 patients. The second external dataset was the UCLA dataset, which was published by [29]. This dataset consisted of 346 patients who were treated between 2011 and 2019. The selection of individuals for the cohorts was based on three specific inclusion criteria. Firstly, they had to have a confirmed diagnosis of AIS. Secondly, they must have undergone a pretreatment MRI scan that included DWI, FLAIR, and ADC series, which were obtained without any motion degradation. Lastly, they needed to have a known time of stroke symptom onset (TSS) within 24 hours of the MRI scan.

Figure 2. Class distributions of clinical defined cut-off time points (3 hours, 4.5 hours, 6 hours) with training and validation dataset



The distribution of the clinical defined cut-off time points is indicated in Figure 2. These time points, which are clinically meaningful, were defined as labels based on the training and validation sets, with the specific values of 3 hours, 4.5 hours, and 6 hours. Since the initial study design used 4.5 hours as the cutoff for training and test sets, the proportion of symptom onset within 3 hours and 6 hours is imbalanced.

Figure 3. Distributions of patients in relation to time from symptom onset with training, internal and external validation sets

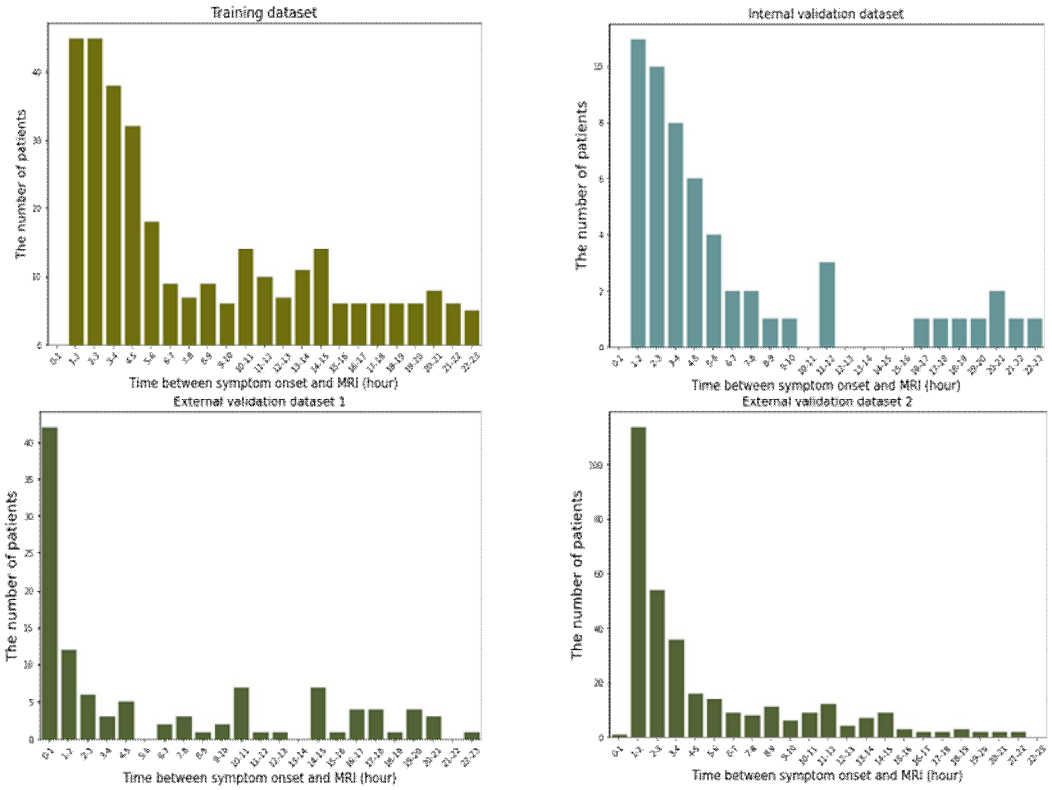


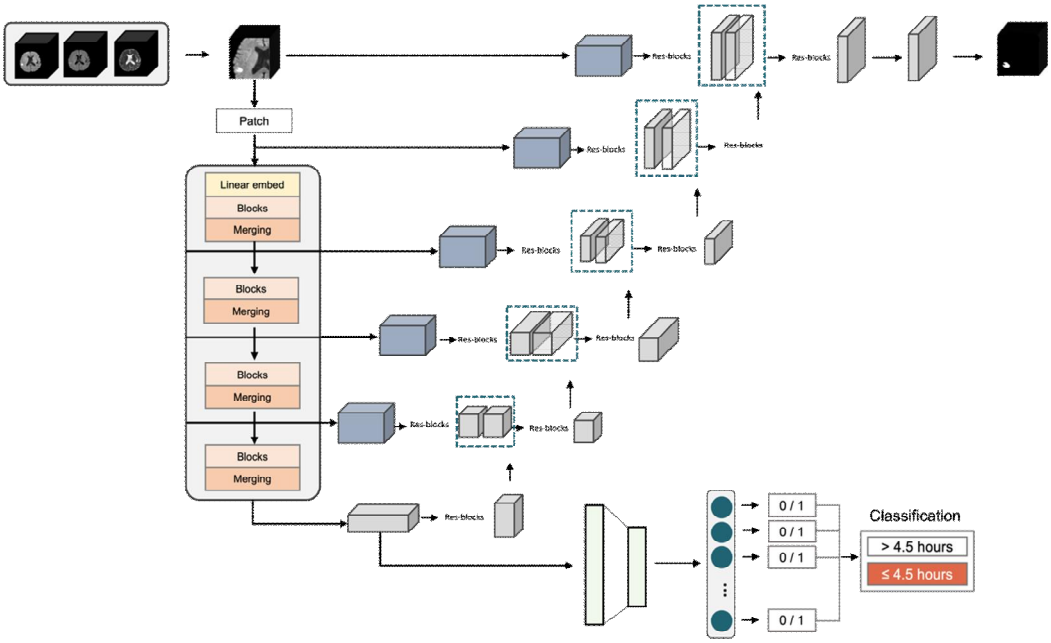
Figure 3 illustrates the distributions of time from symptom onset to MRI in hours across different datasets: the training dataset, internal validation dataset, external validation dataset 1, and external validation dataset 2. The y-axis represents the number of patients, while the x-axis represents the time between symptom onset and MRI in hours. The training dataset exhibits cases in all hour ranges, with the highest number of patients observed in the 1-2- and 2-3 hour ranges, providing comprehensive coverage of patients with varying onset times. Similarly, the internal validation dataset shows a similar pattern, with a greater number of cases in the 1-2 hour range and some hour ranges lacking cases. In contrast, the external validation dataset 1 displays a distinct distribution, concentrating cases primarily in the 0-1 hour range, while the remaining hour ranges have sparser distributions. Finally, the external validation dataset 2 demonstrates a distribution pattern akin to the training and internal validation datasets, with the highest number of cases found in the 1-2 hour range.

These findings provide valuable insights into the temporal distribution of patients' time from symptom onset to MRI across different datasets, contributing to our understanding of the data used in training the mode and its overall performance.

## 2. Model Architecture

We propose a model that simultaneously predicts the onset time of stroke and performs lesion segmentation using MTL for the clinically defined cutoff time points of 3, 4.5, and 6 hours from symptom onset. In the proposed model, a auxiliary classifier is added to the bottleneck layer of the 3D patch based Swin UNETR. Furthermore, a voting ensemble of patch-level classifiers is employed for patient-level classification.

Figure 4. Overview of the proposed multi-task learning network with classification and segmentation task




## 2.1. Data Preprocessing

As a first preprocessing step, FLAIR images were coregistered to the ADC maps using Statistical Parametric Mapping 12 software (SPM 12; Wellcome Trust Centre for Neuroimaging, University College of London, United Kingdom) [45]. To remove the skull, HD-BET was applied to the FLAIR images for brain extraction [46]. Taking into consideration that the images had different voxel spacing for each patient, all the images were resampled to a voxel size of 1mm x 1mm x 3mm from the original ROIs.

### 2.2. Proposed model

To perform infarct lesion segmentation and AIS onset identification, we utilized the 3D patch based Swin UNETR. We added an auxiliary classifier to the bottleneck layer of the model, and the proposed model is illustrated in Figure 1. Swin UNETR has a U-shaped network [47] design in which the extracted feature representations of the encoder are used in the decoder via skip connections at each resolution. To perform patient-level classification, a voting ensemble on the patch level classifiers was performed. A voting ensemble of patch-level classifiers is performed at the patient level for classification. The auxiliary classifiers count and vote for onset identification based on the presence of infarction in patches using the predicted mask.

### 2.3. Multi-task Loss function

For the classification task, class imbalance is a common challenge in medical image classification. For example, the number of 3- and 6-hour windows of our dataset is roughly twice. To account for this, we use a modified weighted focal loss [48] as the classification loss function:

$$L_{cls}(p_{cls}, y_{cls}) = -w_m(1 - p_{cls})^{\gamma} y_{cls} log(p_{cls}) - w_n p_{cls}^{\gamma}(1 - y_{cls})^{\gamma} log(1 - p_{cls}), \quad (1)$$

where $p_{cls}$ and $y_{cls}$ are the predicted volume classification probability from the proposed network and the ground truth class of this volume ($y_{cls}$ = 0 for $\geq$ cut-off time window and $y_{cls}$ = 1 for $\leq$ onset time window). The focusing parameter $\gamma$ is set as 2. $w_m$ and $w_n$ are weights for imbalance classes defined as:

$$w_m = \frac{N_n}{N_m + N_n}, \qquad w_n = \frac{N_m}{N_m + N_n}, \quad (2)$$

where $N_n$ and $N_m$ are the numbers that indicate whether they are included in the cut-off time window or not, respectively.

For the segmentation task, a segmentation loss based on the Dice coefficient.

$$L_{seg}(P_{seg}, Y_{seg}) = 1 - \frac{2P_{seg}Y_{seg} + 1}{P_{seg} + Y_{seg} + 1}, \quad (3)$$

where Lseg is the segmentation loss, Pseg and Yseg denote the predicted segmentation map from the proposed network and the infarct lesion of AIS.

In our method, classification loss and segmentation loss are linear combined as a multi-task loss by a hyperparameter λ. The multi-task loss is defined as:

$$L_{joint} = \lambda L_{cls} + (1 - \lambda)L_{seg}, \quad (4)$$

## 3. Experiments

### 3.1. Data augmentation

Data augmentation techniques included random per-channel intensity shifts within the range of (-0.1, 0.1), and random intensity scaling within the range of (0.9, 1.1) for input image channels. A random axis mirror flip was applied with a probability of 0.5 for all three axes. Input images were normalized to have zero mean and unit standard deviation based on non-zero voxels.

### 3.2. Implementation details

Our model is implemented using PyTorch and MONAI and trained with 8 NVIDIA V100 GPUs. An overview of the shifted windowing mechanism was implemented with the same parameters as described in the Swin UNETR [49]. During training, random patches of size 128 × 128 × 128 were cropped from the 3D image volumes. We chose AdamW [50] as an optimizer with a learning rate of 1e-4. Each GPU had a batch size of 1. The models were trained for 300 epochs with a linear warmup and a cosine annealing learning rate scheduler. During inference, a sliding window approach was used with a 0.7 overlap between neighboring voxels.

### 3.3. Statistical analysis

To compare the performance of stroke onset classification, various metrics of deep learning methods were analyzed, including sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), and accuracy in predicting the recommended time for thrombolysis (within 3 hours, 4.5 hours, and 6 hours). And we employed receiver operating characteristic (ROC) analysis, including determinations of areas under the ROC curves (AUCs) for the quantitative evaluation of stroke onset classification. Also, to evaluate the performance of the infarct segmentation results of the model, the dice similarity coefficient (DSC) and intersection over union (IOU) metrics were used. Furthermore, gradient-weighted class activation map (Grad-CAM) [51] was employed during training to assess whether the model paid attention to the infarct lesion.

## 4. Results

### 4.1. Classification results with several models

Table 2. Model performance comparison with several models (3D DenseNet, 3D patched DenseNet, ours) for classification task

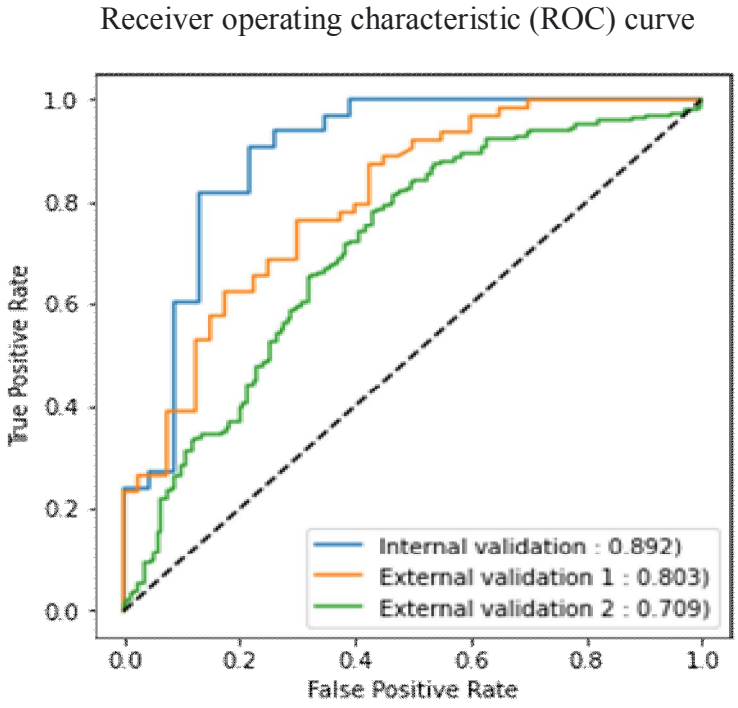
| Model | Onset window | Sensitivity | Specificity | PPV | NPV | Accuracy | AUC |
|---|---|---|---|---|---|---|---|
| 3D DenseNet | 3 hours | 0.667 | 0.478 | 0.647 | 0.500 | 0.589 | 0.617 |
| 3D patched DenseNet | | 0.435 | 0.758 | 0.556 | 0.658 | 0.625 | 0.638 |
| Ours | | 0.696 | 0.576 | 0.533 | 0.731 | 0.625 | **0.690** |
| 3D DenseNet | 4.5 hours | 0.667 | 0.565 | 0.688 | 0.542 | 0.625 | 0.632 |
| 3D patched DenseNet | | 0.424 | 0.913 | 0.875 | 0.525 | 0.625 | 0.769 |
| Ours | | 0.879 | 0.783 | 0.853 | 0.818 | 0.839 | **0.895** |
| 3D DenseNet | 6 hours | 0.436 | 0.882 | 0.895 | 0.405 | 0.571 | 0.691 |
| 3D patched DenseNet | | 0.667 | 0.882 | 0.929 | 0.536 | 0.732 | 0.799 |
| Ours | | 0.897 | 0.647 | 0.854 | 0.733 | 0.821 | **0.848** |

Table 2 is a comparison table of the performance of several models in terms of classification results. The performance of our model in predicting the onset time of stroke was evaluated by comparing it with two other 3D deep learning models, namely 3D DenseNet [52] and 3D patched DenseNet. The models were trained to predict the onset time of stroke events at 3 hours, 4.5 hours, and 6 hours, respectively. The performance of the models was assessed using several performance metrics, including AUC, accuracy, sensitivity, specificity, PPV, and NPV.

Our model achieved superior performance compared to the other models, with AUC values of 0.690, 0.895, and 0.848 and corresponding accuracy values of 0.625, 0.839, and 0.821 for the onset times of 3 hours, 4.5 hours, and 6 hours, respectively. Furthermore, our model demonstrated the highest sensitivity with values of 0.576, 0.783, and 0.647, respectively. While the specificity values were relatively lower, our model showed respectable performance in terms of PPV and outstanding performance in terms of NPV. Overall, our model exhibited robust and superior performance in predicting the onset time of stroke events.

### 4.2. Classification results with external validations

Figure 5. A result of the ROC curve with AUROC score for internal and two external validations



The performance of our model in classifying the stroke onset window was comprehensively evaluated using internal and two external validation datasets. Figure 5 illustrates the ROC curve for internal validation, external validation 1, and external validation 2. The AUC values calculated for internal validation, external validation 1, and external validation 2 were found to be 0.892, 0.803, and 0.709, respectively. Notably, the model exhibited superior discriminatory performance in internal and external validation 1 compared to external validation 2.

Table 3. Model performance of classification task within 4.5-hour window for internal and external validation sets

| Model | Onset window | Sensitivity | Specificity | PPV | NPV | Accuracy | AUC |
|---|---|---|---|---|---|---|---|
| Internal validation | 4.5 hours | 0.879 | 0.783 | 0.853 | 0.818 | 0.839 | 0.895 |
| External validation 1 | | 0.750 | 0.700 | 0.80 | 0.636 | 0.731 | 0.803 |
| External validation 2 | | 0.744 | 0.58 | 0.676 | 0.658 | 0.668 | 0.709 |

The model also exhibited consistent and respectable accuracy across the different datasets, with accuracy values of 0.839, 0.731, and 0.668 for internal validation, external validation 1, and external validation 2, respectively. These results emphasize the reliability and reproducibility of our model's accuracy in predicting the stroke onset window. Furthermore, our model achieved a balanced performance in both sensitivity and specificity, important indicators of diagnostic performance, in both the internal and external validation datasets. For internal validation, the model achieved a sensitivity value of 0.879 and a specificity value of 0.783. Similarly, in external validation 1, the sensitivity value was 0.750, and the specificity value was 0.700. In external validation 2, the model demonstrated a sensitivity value of 0.744 and a specificity value of 0.58. These findings underscore the robustness of

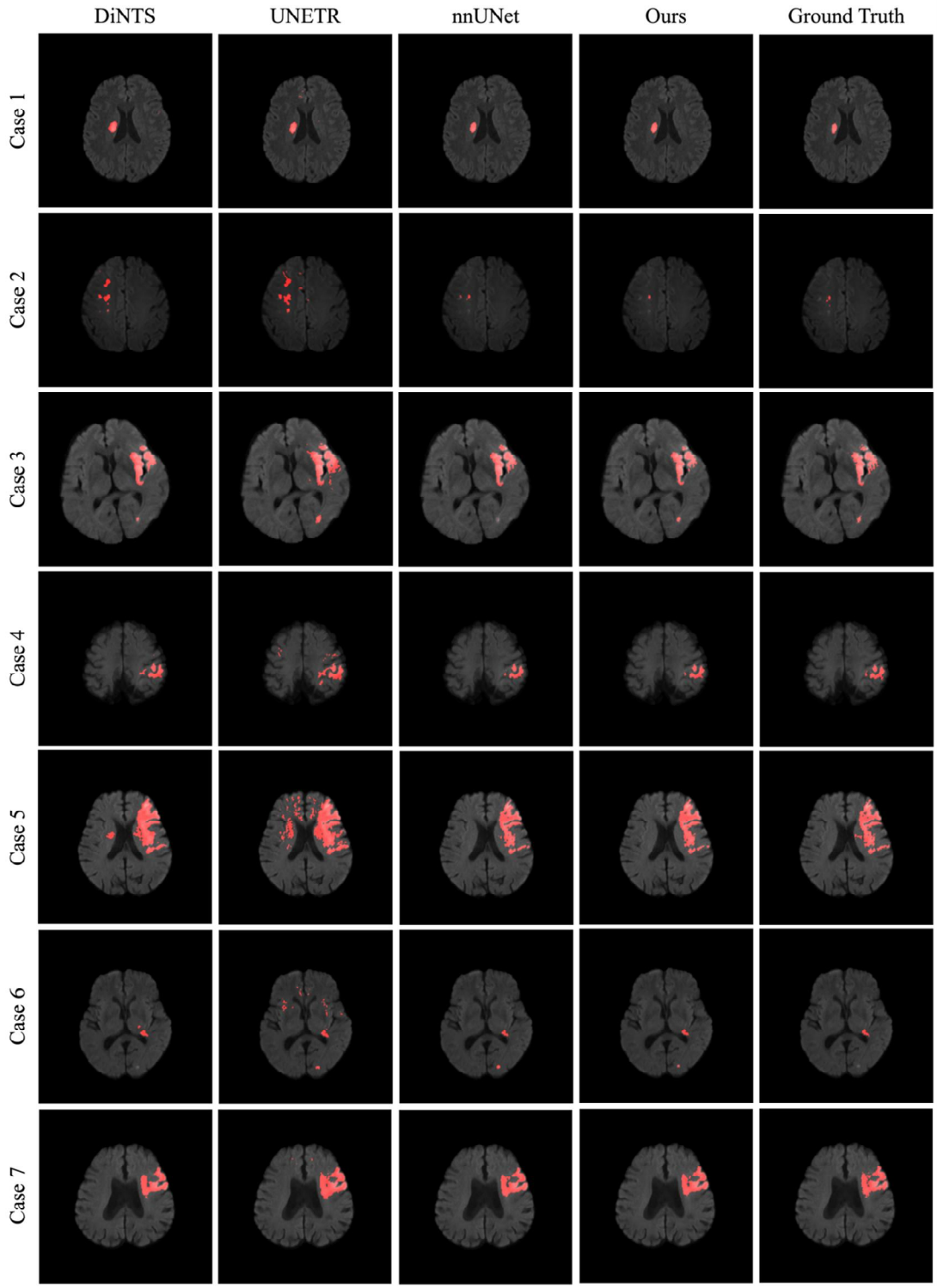
our model in predicting and accurately identifying whether the stroke onset time falls within the 4.5-hour treating window across multiple validation datasets, highlighting its potential for assisting in timely decision-making in clinical settings.

### 4.3. Segmentation results with several models

Figure 6 shows the visualization of the results of ground truth infarct lesion and predictions made by various models based on internal validation. The compared models include DiNTS [53], UNETR [40], nnUNet [35], and our model, all of which have shown relatively good performance in segmentation tasks. Using these models, the prediction results of randomly selected case 7 were compared. Overall, it can be observed that both the DiNTS and UNETR models have made predictions that are excessively different from the ground truth in most of the images. In particular, in images such as case 2, 4, 5, and 6, it can be observed that UNETR has poor prediction performance compared to other models. On the other hand, nnUNet and our model demonstrated superior prediction performance compared to the ground truth in most of the images.

Figure 6. Model performance with several models (DiNTs, UNETR, nnUNet, ours) with ground truth for segmentation task

Figure 6. Model performance with several models (DiNTs, UNETR, nnUNet, ours) with ground truth for segmentation task



### 4.4. Evaluation of segmentation results with internal validation

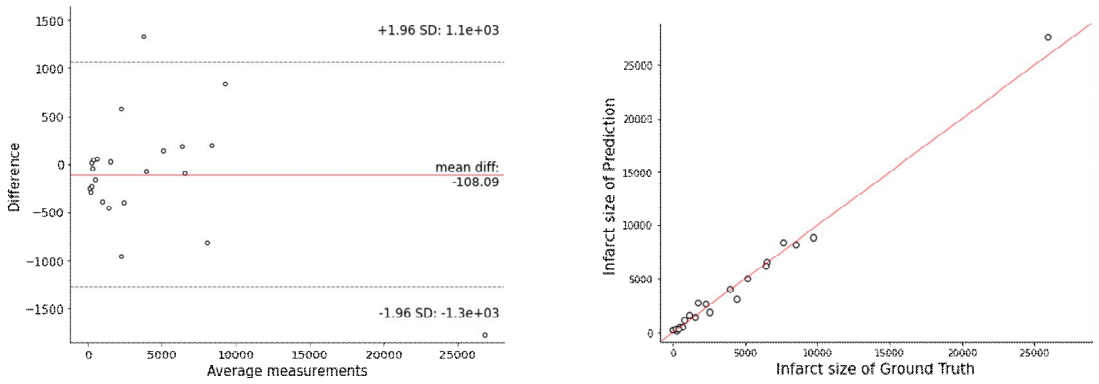
Table 4. Evaluation of segmentation results with several models (DINTS, UNETR, nnUNET, ours) for internal validation

| | DINTS | UNETR | nnUNet | Ours |
|---|---|---|---|---|
| IOU | 0.669 | 0.652 | 0.656 | 0.658 |
| DSC | 0.684 | 0.690 | 0.720 | 0.722 |

In Table 4, the evaluation results of segmentation using internal validation are presented. Since the external validation sets did not have an infarct mask, the metrics for the models were compared only for internal validation. The evaluation metrics used were IOU and DSC. The performance of four models (DINTS, UNETR, nnUNET, ours), which have shown good performance in recent studies, was compared. The IOU results for DINTS, UNETR, nnUNET, and ours were 0.669, 0.652, 0.656, and 0.658, respectively, with ours demonstrating the highest performance. Similarly, for DSC, the results were 0.684, 0.690, 0.720, and 0.722, with ours again achieving the best performance.

4.5. Evaluation of segmentation results with internal validation set

Figure 7. Bland-Altman plot and correlation with segmentation results of internal validation set



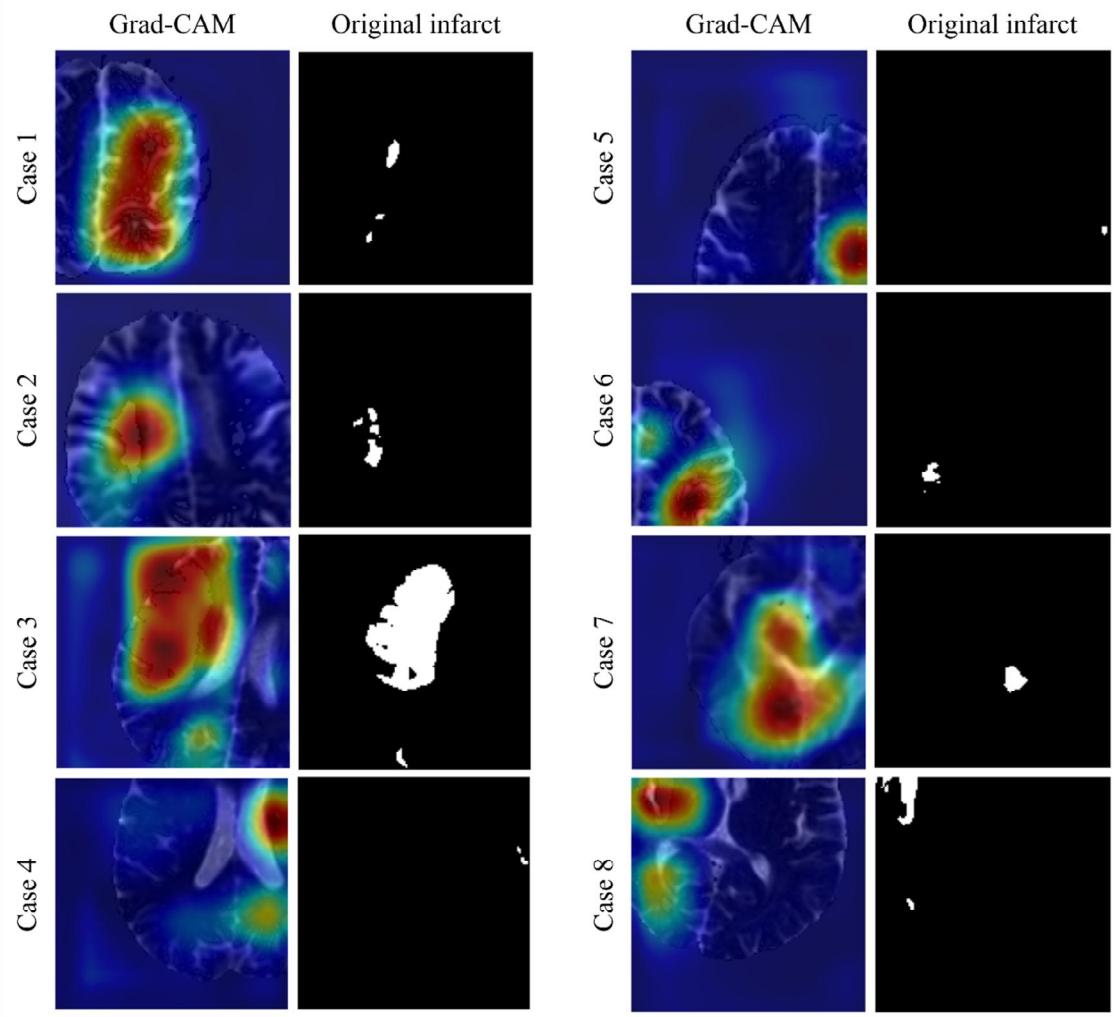

(a) (b)

We employed the Bland-Altman plot, a widely recognized method as a means to assess the agreement between our proposed model's predicted AIS onset time and the corresponding ground truth values. The Bland-Altman plot, as shown in Figure 7(a), presents a graphical representation of the difference between the predicted and ground truth values on the y-axis, while the x-axis represents the average of the two values. The plot allows us to assess the presence of any systematic bias or proportional differences between the predicted and ground truth values. The results demonstrated a general alignment between the predicted and ground truth values, indicating a favorable level of agreement. The two exceptional cases exhibited larger differences between the predicted and ground truth values, suggesting the presence of potential outliers or areas where further investigation may be warranted.

Figure 7(b) showcases a graphical representation of the correlation between the segmentation results obtained from the internal validation set. In this representation, the y-axis illustrates the infarct size as predicted by our model, while the x-axis corresponds to the infarct size as determined by the ground truth. The results demonstrated a strong linear relationship between the predicted infarct sizes and the corresponding ground truth values. This alignment of the correlation plot provides additional evidence supporting the reliability and accuracy of our model.

4.6. Grad-CAM with internal validation set

Figure 8. Grad-CAM results of prediction with our model and original infarction mask




The results for examining the explainability of our model's attention to infarction during training are described in Figure 7. The images were randomly selected from 8 different cases, each with infarctions of varying sizes. For each case, the left image represents the result of Grad-CAM, while the right image shows the original infarction image. The results indicate that the model focused on infarctions in the images while performing onset time prediction with Grad-CAM from its encoder.

Figure 9. Grad-CAM results of prediction using our model with a few bad false positive and false negative cases

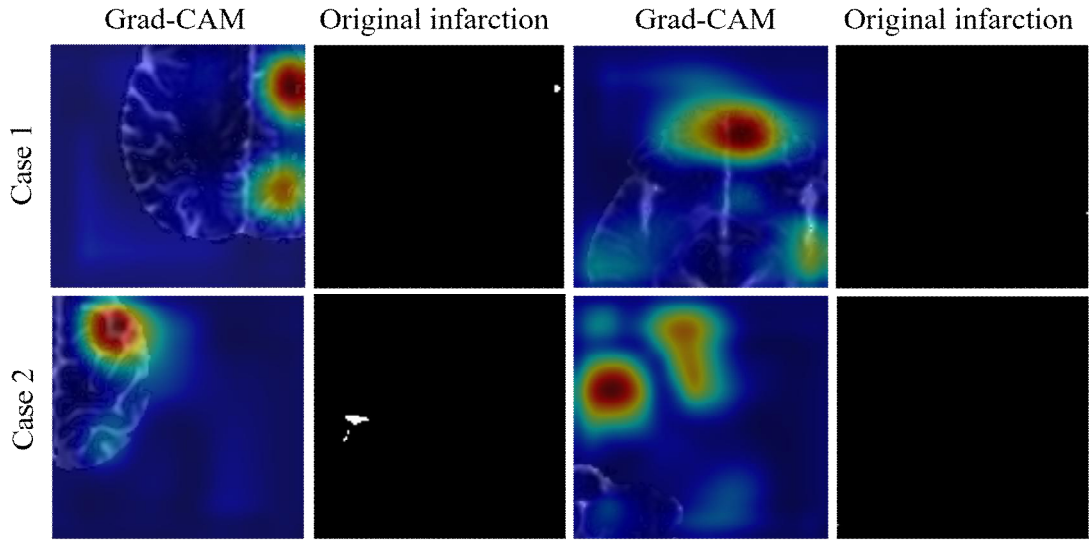



Figure 9 displays the Grad-CAM results of false positive and false negative cases, representing instances where our model exhibited suboptimal performance. The training process utilizes 3D patches with random cropping based on their center, and we hypothesize that there may have contributed to an imbalance between patches with infarct and those without infarct. This imbalance could have potentially affected the model's ability to generalize and accurately classify instances in which infarct is present. The presence of both false positives and false negatives highlights the need for further investigation and potential adjustments in the future.

## Discussion

This study demonstrated the usefulness of an automated system developed MTL with deep learning methods in identifying the tissue age of patients with acute stroke within 4.5 hours of symptom onset. The AIS onset prediction and infarction segmentation results using DWI and FLAIR mismatch were evaluated both quantitatively and qualitatively.

*External validation results with classification task*

The results were verified not only through internal validation but also across diverse datasets. Two external validation sets were further utilized for inference in order to evaluate the robustness and generalizability of our proposed model. The classification result for onset time of 4.5 hours revealed an AUC of 0.803 in external validation set 1 and 0.709 in external validation set 2. External validation set 1 presented a respectable performance despite the differences in demographic characteristics, such as a higher proportion of female patients and delayed onset to MRI [54]. In contrast, the significant drop in AUC for external validation set 2 can be attributed to various factors, including differences in demographic characteristics such as ethnicity and sex, as well as unclear criteria for onset time calculation [55, 56]. Moreover, the use of different MRI protocols and variability in image quality in external validation set 2 may have largely affected the performance of our model.

In addition, we need to consider the potential limitations in our study regarding the prediction of the onset of AIS. Our model has inherent limitations in its architecture. For instance, the brain in the MR images were isolated using foreground cropping, and the patches fed into the model were randomly cropped based on a voxel size of 128 x 128 x 128 from the foregrounded MR images. We hypothesize that, since the size of the brain and the infarct size and location varies for each patient, images without infarction may have been included in the training process. Given that the classification task of onset prediction is based on infarction, this can potentially lead to false negatives (FN) and false positives (FP).

*Internal validation results with segmentation task*

Since there was no external validation available for validating the segmentation predictions, we examined the results using the internal validation set. Among several models, the DiNTS and UNETR models exhibited the least favorable performance, as they tended to over-segment the lesions compared to the actual infarct. On the other hand, the nnUNet and our proposed model demonstrated satisfactory performance, accurately predicting segmentations for infarcts of various sizes. We also validated the performance of these four models using metrics such as IOU (Intersection over Union) and Dice score. The metrics results aligned with the visualization of the segmentation outcomes, generally indicating that nnUNet and our proposed model performed well.

Additionally, for the Bland-Altman plot evaluating the agreement between the segmentation results and the ground truth values, most of the results showed general agreement, excluding rare cases where there were significant differences between the actual values and the predicted values. It was observed that the cases with large differences corresponded to situations where the actual infarct size was exceptionally large.

*Grad-CAM results with our model*

To further validate the interpretability of our model, we visualized the activation map of the last layer of the encoder. The visualization demonstrated that our model predicts infarction with attention on significant features in almost all images, providing further evidence of its high accuracy.

Despite the excellent performance of our proposed model, it still has its limitations. First, only acute ischemic stroke was considered as inclusion criteria, and further studies are required to evaluate different types of strokes. Second, the external validation datasets only

contained onset labels, and no infarction masks were available for segmentation evaluation, and consequently, further studies with the use of datasets with segmentation masks are required to evaluate our model's segmentation accuracy. Finally, external validation set 2 was a blind dataset, and the analysis of the exact cause of the drop in AUC was difficult to determine.

## Conclusion

In conclusion, our proposed model utilizes multi-task learning to perform infarction segmentation and classification of onset time. Our model demonstrated its robustness and enhanced performance through internal and external datasets. While limitations and the need for further study must be acknowledged, we anticipate that our model will have a positive impact in clinical setting in the near future. To facilitate the translation of our model into real-world clinical practice, further research should address its limitations and extend its application to a broader range of stroke types and patient populations.

**References**


1. Virani SS, Alonso A, Benjamin EJ, Bittencourt MS, Callaway CW, Carson AP, et al. Heart disease and stroke statistics—2020 update: a report from the American Heart Association. *Circulation* 2020;141:e139-e596
2. Feigin VL, Brainin M, Norrving B, Martins S, Sacco RL, Hacke W, et al. World Stroke Organization (WSO): global stroke fact sheet 2022. *International Journal of Stroke* 2022;17:18-29
3. Khatri P, Abruzzo T, Yeatts S, Nichols C, Broderick J, Tomsick T. Good clinical outcome after ischemic stroke with successful revascularization is time-dependent. *Neurology* 2009;73:1066-1072
4. Greshman G, Alexander D. National Institute of Neurological Disorders and Stroke rt-PA Stroke Study Group. *Tissue plasogen activator for acute ischemic stroke* 1997;28:1522-1526
5. Powers WJ, Derdeyn CP, Biller J, Coffey CS, Hoh BL, Jauch EC, et al. 2015 American Heart Association/American Stroke Association focused update of the 2013 guidelines for the early management of patients with acute ischemic stroke regarding endovascular treatment: a guideline for healthcare professionals from the American Heart Association/American Stroke Association. *Stroke* 2015;46:3020-3035
6. Powers WJ, Rabinstein AA, Ackerson T, Adeoye OM, Bambakidis NC, Becker K, et al. Guidelines for the early management of patients with acute ischemic stroke: 2019 update to the 2018 guidelines for the early management of acute ischemic stroke: a guideline for healthcare professionals from the American Heart Association/American Stroke Association. *Stroke* 2019;50:e344-e418
7. Disorders NIoN, Group Sr-PSS. Tissue plasminogen activator for acute ischemic stroke. *New England Journal of Medicine* 1995;333:1581-1588
8. Grossman AW, Broderick JP. Advances and challenges in treatment and prevention of ischemic stroke. *Annals of neurology* 2013;74:363-372
9. Nadeau JO, Fang J, Kapral MK, Silver FL, Hill MD. Outcome after stroke upon awakening. *Canadian journal of neurological sciences* 2005;32:232-236

10. George MG, Tong X, McGruder H, Yoon P, Rosamond W, Winquist A, et al. Paul coverdell national acute stroke registry surveillance—four states, 2005–2007. *Morbidity and Mortality Weekly Report: Surveillance Summaries* 2009;58:1-23
11. Serena J, Dávalos A, Segura T, Mostacero E, Castillo J. Stroke on awakening: looking for a more rational management. *Cerebrovascular diseases* 2003;16:128-133
12. Thomalla G, Cheng B, Ebinger M, Hao Q, Tourdias T, Wu O, et al. DWI-FLAIR mismatch for the identification of patients with acute ischaemic stroke within 4· 5 h of symptom onset (PRE-FLAIR): a multicentre observational study. *The Lancet Neurology* 2011;10:978-986
13. Fisher M, Albers GW. Advanced imaging to extend the therapeutic time window of acute ischemic stroke. *Annals of neurology* 2013;73:4-9
14. Wu O, Schwamm LH, Sorensen AG. Imaging stroke patients with unclear onset times. *Neuroimaging Clinics* 2011;21:327-344
15. Aoki J, Kimura K, Iguchi Y, Shibazaki K, Sakai K, Iwanaga T. FLAIR can estimate the onset time in acute ischemic stroke patients. *Journal of the neurological sciences* 2010;293:39-44
16. Moseley M, Kucharczyk J, Mintorovitch J, Cohen Y, Kurhanewicz J, Derugin N, et al. Diffusion-weighted MR imaging of acute stroke: correlation with T2-weighted and magnetic susceptibility-enhanced MR imaging in cats. *American Journal of Neuroradiology* 1990;11:423-429
17. Mintorovitch J, Moseley M, Chileuitt L, Shimizu H, Cohen Y, Weinstein P. Comparison of diffusion□and T2□weighted MRI for the early detection of cerebral ischemia and reperfusion in rats. *Magnetic resonance in medicine* 1991;18:39-50
18. Hoehn□berlar M, Eis M, Back T, Kohno K, Yamashita K. Changes of relaxation times (T1, T2) and apparent diffusion coefficient after permanent middle cerebral artery occlusion in the rat: temporal evolution, regional extent, and comparison with histology. *Magnetic resonance in medicine* 1995;34:824-834
19. Venkatesan R, Lin W, Gurleyik K, He Y, Paczynski R, Powers W, et al. Absolute measurements of water content using magnetic resonance imaging: preliminary findings in an in vivo focal ischemic rat model. *Magnetic Resonance in Medicine:*

*An Official Journal of the International Society for Magnetic Resonance in Medicine* 2000;43:146-150

20. Ebinger M, Galinovic I, Rozanski M, Brunecker P, Endres M, Fiebach JB. Fluid-attenuated inversion recovery evolution within 12 hours from stroke onset: a reliable tissue clock? *Stroke* 2010;41:250-255
21. Petkova M, Rodrigo S, Lamy C, Oppenheim G, Touzé E, Mas J-L, et al. MR imaging helps predict time from symptom onset in patients with acute stroke: implications for patients with unknown onset time. *Radiology* 2010;257:782-792
22. Emeriau S, Serre I, Toubas O, Pombourcq F, Oppenheim C, Pierot L. Can diffusion-weighted imaging–fluid-attenuated inversion recovery mismatch (positive diffusion-weighted imaging/negative fluid-attenuated inversion recovery) at 3 Tesla identify patients with stroke at< 4.5 Hours? *Stroke* 2013;44:1647-1651
23. Grosse-Dresselhaus F, Galinovic I, Villringer K, Audebert HJ, Fiebach JB. Difficulty of MRI based identification of lesion age by acute infra-tentorial ischemic stroke. *PLoS One* 2014;9:e92868
24. Thomalla G, Rossbach P, Rosenkranz M, Siemonsen S, Krützelmann A, Fiehler J, et al. Negative fluid□attenuated inversion recovery imaging identifies acute ischemic stroke at 3 hours or less. *Annals of Neurology: Official Journal of the American Neurological Association and the Child Neurology Society* 2009;65:724-732
25. Kim BJ, Kim Y-H, Kim Y-J, Ahn SH, Lee DH, Kwon SU, et al. Color-coded fluid-attenuated inversion recovery images improve inter-rater reliability of fluid-attenuated inversion recovery signal changes within acute diffusion-weighted image lesions. *Stroke* 2014;45:2801-2804
26. Lee H, Lee E-J, Ham S, Lee H-B, Lee JS, Kwon SU, et al. Machine learning approach to identify stroke within 4.5 hours. *Stroke* 2020;51:860-866
27. Zhu H, Jiang L, Zhang H, Luo L, Chen Y, Chen Y. An automatic machine learning approach for ischemic stroke onset time identification based on DWI and FLAIR imaging. *NeuroImage: Clinical* 2021;31:102744
28. Ho KC, Speier W, El-Saden S, Arnold CW. Classifying acute ischemic stroke onset time using deep imaging features. In:*AMIA Annual Symposium Proceedings*:

American Medical Informatics Association, 2017; 892

29. Polson JS, Zhang H, Nael K, Salamon N, Yoo BY, El□Saden S, et al. Identifying acute ischemic stroke patients within the thrombolytic treatment window using deep learning. *Journal of Neuroimaging* 2022;32:1153-1160

30. Ruder S. An overview of multi-task learning in deep neural networks. *arXiv preprint arXiv:1706.05098* 2017

31. Wang P, Patel VM, Hacihaliloglu I. Simultaneous segmentation and classification of bone surfaces from ultrasound using a multi-feature guided CNN. In:*Medical Image Computing and Computer Assisted Intervention–MICCAI 2018: 21st International Conference, Granada, Spain, September 16-20, 2018, Proceedings, Part IV 11*: Springer, 2018; 134-142

32. Song L, Lin J, Wang ZJ, Wang H. An end-to-end multi-task deep learning framework for skin lesion analysis. *IEEE journal of biomedical and health informatics* 2020;24:2912-2921

33. Kamnitsas K, Bai W, Ferrante E, McDonagh S, Sinclair M, Pawlowski N, et al. Ensembles of multiple models and architectures for robust brain tumour segmentation. In:*Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries: Third International Workshop, BrainLes 2017, Held in Conjunction with MICCAI 2017, Quebec City, QC, Canada, September 14, 2017, Revised Selected Papers 3*: Springer, 2018; 450-462

34. Zhou C, Chen S, Ding C, Tao D. Learning contextual and attentive information for brain tumor segmentation. In:*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, 2019; 497-507

35. Isensee F, Jäger PF, Full PM, Vollmuth P, Maier-Hein KH. nnU-Net for brain tumor segmentation. In:*Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries: 6th International Workshop, BrainLes 2020, Held in Conjunction with MICCAI 2020, Lima, Peru, October 4, 2020, Revised Selected Papers, Part II 6*: Springer, 2021; 118-132

36. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. *Advances in neural information processing systems* 2017;30
37. Devlin J, Chang M-W, Lee K, Toutanova K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* 2018
38. Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929* 2020
39. Raghu M, Unterthiner T, Kornblith S, Zhang C, Dosovitskiy A. Do vision transformers see like convolutional neural networks? *Advances in Neural Information Processing Systems* 2021;34:12116-12128
40. Hatamizadeh A, Tang Y, Nath V, Yang D, Myronenko A, Landman B, et al. Unetr: Transformers for 3d medical image segmentation. In:*Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 2022; 574-584
41. Liu Z, Lin Y, Cao Y, Hu H, Wei Y, Zhang Z, et al. Swin transformer: Hierarchical vision transformer using shifted windows. In:*Proceedings of the IEEE/CVF international conference on computer vision*, 2021; 10012-10022
42. Liu Z, Ning J, Cao Y, Wei Y, Zhang Z, Lin S, et al. Video swin transformer. In:*Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022; 3202-3211
43. Hatamizadeh A, Nath V, Tang Y, Yang D, Roth HR, Xu D. Swin unetr: Swin transformers for semantic segmentation of brain tumors in mri images. In:*International MICCAI Brainlesion Workshop*: Springer, 2021; 272-284
44. Wechsler H. Texture analysis—a survey. *Signal Processing* 1980;2:271-282
45. Ashburner J, Friston KJ. Unified segmentation. *Neuroimage* 2005;26:839-851
46. Isensee F, Schell M, Pflueger I, Brugnara G, Bonekamp D, Neuberger U, et al. Automated brain extraction of multisequence MRI using artificial neural networks. *Human brain mapping* 2019;40:4952-4964
47. Ronneberger O, Fischer P, Brox T. U-net: Convolutional networks for biomedical image segmentation. In:*Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany,*

*October 5-9, 2015, Proceedings, Part III 18*: Springer, 2015; 234-241

48. Lin T-Y, Goyal P, Girshick R, He K, Dollár P. Focal loss for dense object detection. In:*Proceedings of the IEEE international conference on computer vision*, 2017; 2980-2988

49. Hatamizadeh A, Nath V, Tang Y, Yang D, Roth HR, Xu D. Swin unetr: Swin transformers for semantic segmentation of brain tumors in mri images. In:*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*: Springer, 2022; 272-284

50. Loshchilov I, Hutter F. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101* 2017

51. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-cam: Visual explanations from deep networks via gradient-based localization. In:*Proceedings of the IEEE international conference on computer vision*, 2017; 618-626

52. Huang G, Liu Z, Van Der Maaten L, Weinberger KQ. Densely connected convolutional networks. In:*Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017; 4700-4708

53. He Y, Yang D, Roth H, Zhao C, Xu D. Dints: Differentiable neural network topology search for 3d medical image segmentation. In:*Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021; 5841-5850

54. Ruigrok AN, Salimi-Khorshidi G, Lai M-C, Baron-Cohen S, Lombardo MV, Tait RJ, et al. A meta-analysis of sex differences in human brain structure. *Neuroscience & Biobehavioral Reviews* 2014;39:34-50

55. Brickman AM, Schupf N, Manly JJ, Luchsinger JA, Andrews H, Tang MX, et al. Brain morphology in older African Americans, Caribbean Hispanics, and whites from northern Manhattan. *Archives of neurology* 2008;65:1053-1061

56. Williams CM, Peyre H, Toro R, Ramus F. Neuroanatomical norms in the UK Biobank: the impact of allometric scaling, sex, and age. *Human Brain Mapping*

2021;42:4623-4642

**Abstract (with Korean)**

Time clock 과 tissue progression 사이의 강한 연관성이 있는 급성 허혈성 뇌졸중의 치료는 onset 시간에 매우 의존적입니다. 그러나 수면 중 뇌졸중의 발생이나 unclear 한 onset 등으로 인해 onset 의 발생 시기를 정확하게 판단하기 어렵기 때문에, 이를 해결하기 위해 DWI 와 FLAIR 영상의 신호 불일치를 이용하여 뇌졸중 후 경과 시간을 추측할 수 있는 관련된 연구들이 진행되어 왔습니다. 본 연구에서는 MRI 의 이미지 바이오 마커를 통해 deep learning 의 multi-task learning (MTL) 방식을 이용하여 뇌졸중 이후의 시간을 결정하는 기술을 개발하고 외부 검증 데이터셋을 통해 성능을 평가하는 것이 목표입니다.

제안된 모델의 입력 이미지는 ADC, FLAIR 및 DWI-b1000 volume 의 조합으로 구성되며, 이들의 3, 4.5 및 6 시간 발병 시간에 대해서 MTL 을 사용하여 stroke 의 onset 시간 예측 및 병변 분할을 동시에 진행하였습니다. 모델의 backbone 은 3D patch-based SwinUNETR 모델로, 이 모델의 bottleneck layer 에 보조 분류기를 추가하였으며 환자 수준 분류를 수행하기 위해 패치 수준 분류기의 voting 앙상블을 수행하였습니다. 보조 분류기는 predicted mask 를 기준으로 한 patch 에서 infarction 존재하는 경우에 대해서 카운팅 및 voting 방식을 통해서 onset identifying 을 도출하였습니다. Classification task 에서, 내부 검증을 통해 각각 3 시간, 4.5 시간, 6 시간을 기준으로 우리의 모델과 3D DenseNet 및 3D patched DenseNet 의 성능을 비교하였습니다. 평가 지표로 AUC, accuracy, specificity 및 sensitivity 를 사용하였습니다. 또한, 2 개의 외부 검증을 이용하여 4.5 시간에 대한 onset classification 성능을 비교하였습니다. Segmentation task 에서, DiNTS, UNETR, nnUNET 그리고 제안된 모델의 성능을 비교하였습니다. 평가 지표로는 IOU 및 Dice coefficient 가 사용되었습니다. 전반적으로 classification 결과는 비교된 모델에 비해 제안된 모델의 성능이 우수하였고 specificity 와 sensitivity 가 유사한 성능을 보였습니다. Infarct segmentation 의 정성적인 결과는 DiNTs, UNETR, nnUNet, our model 중 nnUNet 모델과 우리의 모델이 좋은 성능을 보였습니다. IOU, Dice score 를 통한 정량적 평가 또한 정성적인 결과와 일치된 결과를 보였습니다. 모델을 학습할 때 우리의 모델의 explainability 를 확인하기 위해, 모델의 encoder 로부터 추출한 feature 들을 이용하여 Grad-CAM 을 확인한 결과, 학습 시 모델이 병변에 집중하며 수행한다는 것을 보였습니다. 이를 통해 여러 segmentation 및 classification task 에서 정량적, 정성적인 평가를 모두 보였습니다. MTL 을 적용한 이

모델은 내부 및 두 개의 외부 검증에서 뇌졸중 발병 시간을 식별하는 데 더 나은 성능을 보여 실제 임상 환경에서의 잠재적인 사용 가능성을 나타냅니다. 제안된 모델은 onset 시간을 예측하여 발병 시간이 불명확한 뇌졸중 환자의 적절한 시간에 혈전 용해 요법 치료 수행에 도움을 줄 수 있는 잠재력이 있습니다.