

DETECTING ABNORMAL AREAS ON BRAIN MRI IMAGES WITH SWIN-UNET**Le Minh Loi^{1,2}, Tran Nguyen Minh Thu^{2*}, Ho Quoc An², Pham Nguyen Khang²**¹Can Tho University of Medicine and Pharmacy, ²Can Tho University

ARTICLE INFO		ABSTRACT
Received:	07/4/2024	To identify abnormal areas on brain MRI images, radiologists need to examine many slices from the image set. This research helps automatically suggest abnormal areas of the brain on MRI images. The Unet, ResNet, Swin-Unet models are trained on the Can Tho University of Medicine and Pharmacy Hospital data set combined with the LGG data set to segment images with or without abnormal regions. The model will then suggest the abnormal region through the boundary drawn around it. Experimental results show that, when dividing random data by image, the Swin-Unet model achieves the highest accuracy with 0.88, along with Recall, Precision and F1 Score of 0.96, 0.71, and 0.82 respectively. For determining the location and shape of the abnormal region, Swin-Unet also demonstrated high performance with mIoU reaching 0.89 and mDSC reaching 0.91. When dividing the data by patient, the Swin-Unet model once again showed good performance with Accuracy reaching 0.86, along with Recall of 0.88, Precision of 0.79, F1 Score of 0.83, and for mIoU it achieved 0.84 and mDSC reached 0.89. Research results show that the Swin-Unet model has good results in the problem of detecting abnormal areas on brain MRI images.
Revised:	10/6/2024	
Published:	10/6/2024	
KEYWORDS		
Detecting abnormalities		
Medical image segmentation		
Deep learning		
Transformer		
SwinUnet		

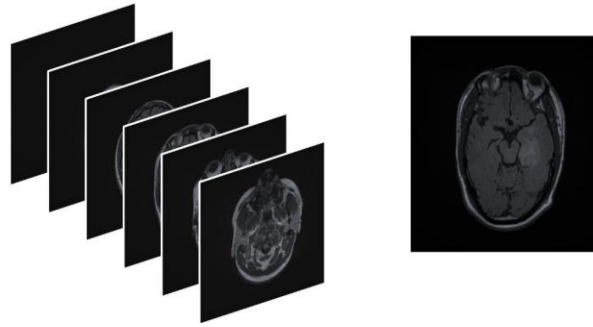
PHÁT HIỆN VÙNG BẤT THƯỜNG TRÊN ẢNH MRI NÃO VỚI MÔ HÌNH SWIN-UNET**Lê Minh Lợi^{1,2}, Trần Nguyễn Minh Thu^{2*}, Hồ Quốc An², Phạm Nguyên Khang²**¹Trường Đại học Y Dược Cần Thơ, ²Trường Đại học Cần Thơ

THÔNG TIN BÀI BÁO		TÓM TẮT
Ngày nhận bài:	07/4/2024	Để xác định vùng bất thường trên ảnh MRI sọ não, bác sĩ chẩn đoán hình ảnh cần khảo sát nhiều lát cắt từ bộ ảnh. Nghiên cứu này giúp tự động phát hiện vùng bất thường của não trên ảnh MRI. Các mô hình Unet, ResNet, Swin-Unet được huấn luyện trên bộ dữ liệu của Bệnh viện Trường Đại học Y Dược Cần Thơ kết hợp bộ dữ liệu LGG để phân đoạn ảnh có hoặc không có vùng bất thường. Sau đó mô hình sẽ đề xuất vùng bất thường thông qua đường biên được vẽ xung quanh. Kết quả thực nghiệm cho thấy, khi chia dữ liệu ngẫu nhiên theo ảnh, mô hình Swin-Unet đạt được độ chính xác cao nhất là 0,88, cùng với Recall, Precision và F1 Score lần lượt là 0,96, 0,71, và 0,82. Đối với việc xác định vị trí và hình dạng của vùng bất thường, Swin-Unet cũng thể hiện hiệu suất cao với mIoU đạt 0,89 và mDSC là 0,91. Khi chia dữ liệu theo bệnh nhân, mô hình Swin-Unet lại một lần nữa thể hiện hiệu suất tốt với độ chính xác (Accuracy) đạt 0,86, cùng với Recall là 0,88, Precision là 0,79, F1 Score là 0,83, còn đối với mIoU đạt 0,84 và mDSC đạt 0,89. Kết quả nghiên cứu cho thấy mô hình Swin-Unet có kết quả tốt trong bài toán phát hiện vùng bất thường trên ảnh MRI não.
Ngày hoàn thiện:	10/6/2024	
Ngày đăng:	10/6/2024	
TỪ KHÓA		
Phát hiện vùng bất thường		
Phân đoạn ảnh y tế		
Học sâu		
Transformer		
Swin-Unet		

DOI: <https://doi.org/10.34238/tnu-jst.10053>* Corresponding author. Email: tnmthu@ctu.edu.vn

1. Giới thiệu

Ảnh X-quang là một ảnh xám đơn lẻ, ảnh MRI não là tập hợp nhiều ảnh xám, mỗi ảnh gọi là lát cắt thể hiện ảnh chụp tại một vị trí cắt ngang của não được thể hiện như Hình 1. Tùy vào cấu hình máy chụp, mỗi bộ ảnh MRI có thể có 16, 32 hoặc 64 lát cắt. Vùng bất thường trên mỗi ảnh có màu sắc, cấu trúc, hình dạng khác với vùng ảnh thông thường. Để khảo sát các dạng vùng bất thường khác nhau (các bệnh khác nhau của não), người ta chụp với các chuỗi xung khác nhau tạo ra các bộ ảnh MRI khác nhau như FLAIR, DWI và STIR.



Hình 1. Vị trí của một cắt MRI trong một bộ ảnh MRI

Ban đầu các phương pháp phát hiện vùng bất thường chủ yếu dựa vào các thuật toán máy học truyền thống [1], [2]. Với sự phát triển của CNN sâu, Unet được đề xuất trong [3], [4] để phân đoạn hình ảnh. Với sự đơn giản và hiệu suất vượt trội của cấu trúc hình chữ U, các biến thể Unet được đề xuất như ResNet [5], Unet ++ [6], DenseUNet [7] và 3dUnet [8].

Gần đây, những thành tựu của mô hình Transformer trong lĩnh vực xử lý ngôn ngữ tự nhiên (NLP) cũng được ứng dụng trong lĩnh vực thị giác máy tính [9]. Trong [10], tác giả đã đề xuất một Transformer tiên phong (ViT) để phân đoạn hình ảnh. ViT đạt kết quả nhỉnh hơn so với các phương pháp dựa trên CNN. Tuy nhiên nhược điểm của ViT là mô hình phải được huấn luyện trước với số rất lớn dữ liệu riêng mới đạt kết quả tốt, trong khi khó có được dữ liệu lớn ảnh y tế được gán nhãn sẵn. Khắc phục nhược điểm này, các biến thể DeiT và Swin Transformer được đề xuất. Cùng với xu hướng đó, mô hình Swin Unet Transformer [11] được đề xuất để phân đoạn ảnh sử dụng kiến trúc chữ U với Swin Transformer làm bộ mã hóa kết nối với bộ giải mã CNN tại các giai đoạn khác nhau thông qua các kết nối. Mô hình được dùng để phân đoạn khối u não 3D đa phương thức với dữ liệu BraTS 2021.

Mô hình Swin Unet [12] là một Transformer thuần của Unet được dùng để phân đoạn ảnh y tế với bộ dữ liệu Synapse multi-organ segmentation dataset (Synapse) và bộ dữ liệu Automated cardiac diagnosis challenge dataset (ACDC). Synapse dataset là bộ dữ liệu gồm 30 bộ ảnh CT bụng với 3779 lát cắt, Swin Unet phân đoạn chính xác 79,13%(DSC $\uparrow$) và 21,55%(HD $\downarrow$). Khi tác giả sử dụng mô hình Swin-Unet để phân đoạn dữ liệu ACDC (ảnh MRI tim) cho độ chính xác 90% (DSC $\uparrow$).

Với sự thành công của Unet và các biến thể trong bài toán phân đoạn ảnh và sự phát triển của Transformer trong lĩnh vực thị giác máy tính. Mô hình Unet dựa trên học sâu CNN, mô hình Swin-Unet dựa trên Transformer được chúng tôi sử dụng để phát hiện vùng bất thường trên ảnh MRI não. Dữ liệu thu thập từ Bệnh viện Trường Đại học Y Dược Cần Thơ kết hợp dữ liệu LGG được tiền xử lý, chuẩn hóa, và được đưa vào thực nghiệm trên các mô hình Unet, ResNet, Swin-Unet nhằm tìm ra mô hình hiệu quả cho bài toán phát hiện vùng bất thường trên ảnh MRI não.

Các nội dung tiếp theo của nghiên cứu được đề cập là mô tả phương pháp nghiên cứu, các kết quả được bàn luận chi tiết để từ đó đưa ra kết luận và hướng phát triển của nghiên cứu.

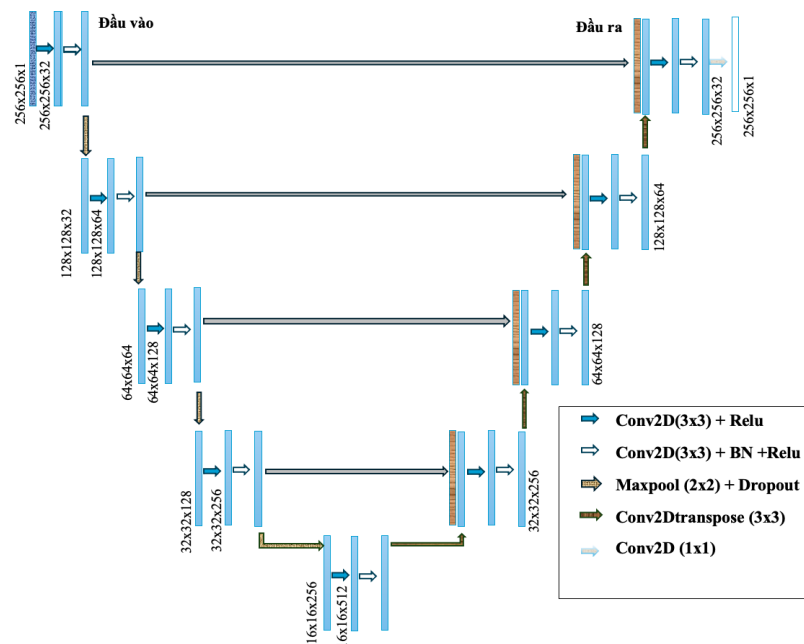
2. Phương pháp nghiên cứu

Mô hình Swin-Unet được áp dụng để phát hiện vùng bất thường trên ảnh MRI não. Bên cạnh đó, các mô hình học sâu đã được phát triển trước đó như Unet, ResUnet cũng được thực nghiệm để so sánh và đánh giá với mô hình SwinUnet. Việc phát hiện vùng bất thường sẽ được đánh giá thông qua các chỉ số của bài toán phân loại hình ảnh như: Độ chính xác (accuracy), độ nhạy (Sensitivity), độ đặc hiệu (specifity) và độ F1-Score. Đối với phân đoạn vùng bất thường sẽ được đánh giá thông qua các chỉ số Mean Intersection over Union (mIoU), Mean Dice Similarity Coefficient (mDSC), Mean Precision, và Mean Recall.

2.1. Mô hình Unet

Mô hình Unet áp dụng để so sánh với mô hình Swin-Unet đề xuất được mô tả trong hình 2. Ảnh đầu vào có kích thước $256 \times 256 \times 1$. Phần giải mã dùng để thu thập ngữ cảnh và các đặc trưng quan trọng từ ảnh thông qua các khối giảm kích thước tương ứng với bộ lọc có số lượng 32, 64, 128 và 256. Trong mỗi khối giảm kích thước sử dụng hai lớp tích chập liên tiếp, mỗi lớp có cùng số lượng bộ lọc với khối đó. Mỗi lớp tích chập theo sau là một hàm kích hoạt ReLU và lớp thứ hai trong mỗi khối cũng được chuẩn hóa theo BatchNormalization. Sau mỗi khối giảm kích thước, sử dụng một lớp MaxPooling2D để giảm kích thước của bản đồ đặc trưng một nửa, và một lớp Dropout để giảm hiện tượng quá mức.

Phần nút cổ chai (Bottleneck): sử dụng một chuỗi gồm 2 lớp tích chập liên tiếp, mỗi lớp đều sử dụng 512 bộ lọc. Theo sau mỗi lớp tích chập là một hàm kích hoạt ReLU. Ở lớp tích chập thứ hai có thêm một lớp chuẩn hóa BatchNormalization.



Hình 2. Kiến trúc mô hình Unet được áp dụng trong thực nghiệm

Phần giải mã (Decoder) phục hồi lại kích thước ban đầu của ảnh và tạo ra bản đồ phân đoạn. Chúng tôi sử dụng bốn khối tăng kích thước, tương ứng với các bộ lọc có số lượng là 256, 128, 64 và 32. Bắt đầu bằng việc tăng kích thước của bản đồ đặc trưng sử dụng lớp Conv2Dtranspose (hay còn gọi là deconvolution) với mục đích tái tạo kích thước không gian gần bằng với bản đồ đặc trưng tương ứng trên phần mã hóa. Sau khi tăng kích thước, bản đồ đặc trưng từ decoder được nối với bản đồ đặc trưng tương ứng từ encoder thông qua phép nối (concatenate). Kết đến là

một lớp Dropout được sử dụng. Cuối cùng, áp dụng một khối tích chập kép, mỗi khối bao gồm hai lớp tích chập với số lượng bộ lọc giảm dần (256, 128, 64 và 32) và hàm kích hoạt ReLU

Sau quá trình giải mã, mô hình cuối cùng sử dụng một lớp tích chập với một bộ lọc và hàm kích hoạt sigmoid để tạo ra bản đồ phân đoạn cuối cùng như Hình 2. Bản đồ này dự đoán xác suất mỗi pixel thuộc về một lớp cụ thể, với kích thước bằng với ảnh đầu vào.

Trong nghiên cứu này, hàm tối ưu hóa Adam với tỉ lệ học là $5e-4$ và epsilon là $1e-5$ được sử dụng. Kích thước lô (batch_size) đặt là 32. Mô hình tự dừng lại sau 150 epoch không có sự giảm về val_loss. Hàm mất mát sử dụng để huấn luyện trong nghiên cứu này là hàm tversky_loss.

2.2. Mô hình Res-Unet

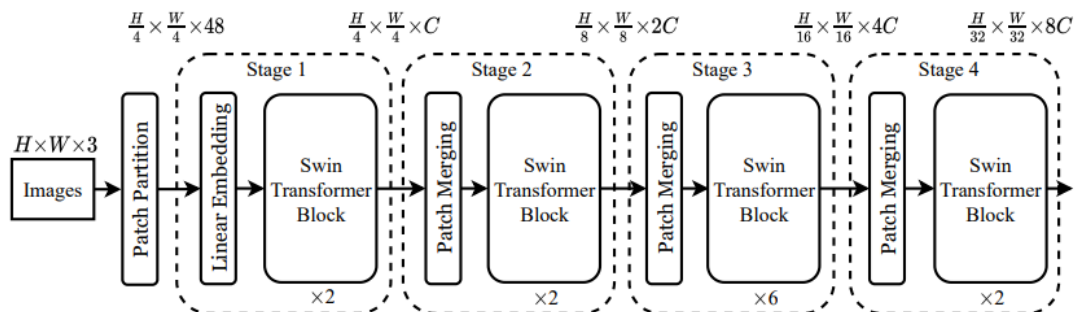
Mô hình ResUnet bắt đầu với lớp đầu vào có kích thước (256,256,1), sau đó triển khai một loạt các khối mã hóa, mỗi khối bao gồm các lớp tích chập với chuẩn hóa hàng loạt và hàm kích hoạt ReLU. Mạng bắt đầu với hai lớp tích chập có 16 bộ lọc 3×3 , sau đó là một kết nối tắt, cũng với 16 bộ lọc nhưng kích thước 1×1 , nhằm cung cấp một dạng kích hoạt ban đầu cho mạng. Tiếp theo, mạng sử dụng ba khối residual trong quá trình mã hóa, mỗi khối tăng gấp đôi số lượng bộ lọc từ 32, 64, đến 128, và áp dụng strides 2 để giảm kích thước của đặc trưng. Mỗi khối này gồm hai lớp tích chập với chuẩn hóa hàng loạt và ReLU, và một kết nối tắt được thực hiện qua một lớp tích chập 1×1 . Phần "cầu nối" của mạng là một khối residual với 256 bộ lọc và strides=2, nối giữa phần mã hóa và giải mã. Trong phần giải mã, mạng thực hiện bốn bước giải mã, mỗi bước sử dụng lớp UpSampling2D để tăng kích thước của đặc trưng và lớp Concatenate để kết hợp các đặc trưng từ phần mã hóa thông qua cơ chế kết nối bỏ qua sau đó là một khối residual, với số lượng bộ lọc giảm dần từ 128, 64, 32 xuống 16.

Cuối cùng, một lớp tích chập với một bộ lọc 1×1 được áp dụng để tạo ra bản đồ phân đoạn cuối cùng, sử dụng hàm kích hoạt "sigmoid" cho phép phân loại từng pixel vào một trong các lớp đích. Kiến trúc ResUNet được thiết kế để tối ưu hóa việc học đặc trưng từ dữ liệu hình ảnh, với khả năng cải thiện đáng kể độ chính xác của phân đoạn ngữ nghĩa so với các mô hình truyền thống, nhờ vào sự kết hợp thông minh giữa các khối residual và cơ chế kết nối bỏ qua của Unet.

Trong nghiên cứu này, hàm tối ưu hóa Adam với tỉ lệ học là $5e-4$ và epsilon là $1e-5$ được sử dụng. Kích thước lô (batch_size) đặt là 32. Ở đây mô hình tự dừng lại sau 150 epoch không có sự giảm về val_loss. Hàm mất mát sử dụng để huấn luyện ở đây là hàm tversky_loss.

2.3. Mô hình Swin-Unet

Mô hình Swin-Unet được cài đặt trong nghiên cứu này là sự kết hợp giữa cấu trúc của Swin Transformer [13] như Hình 3 và kiến trúc mạng Unet. Mạng nhận đầu vào là ảnh có kích thước $256 \times 256 \times 3$ và sử dụng kiến trúc Swin-T cho bộ mã hóa để trích xuất thông tin đặc trưng từ ảnh đầu vào.



Hình 3. Kiến trúc mô hình Swin Transformer phiên bản Swin-T [13]

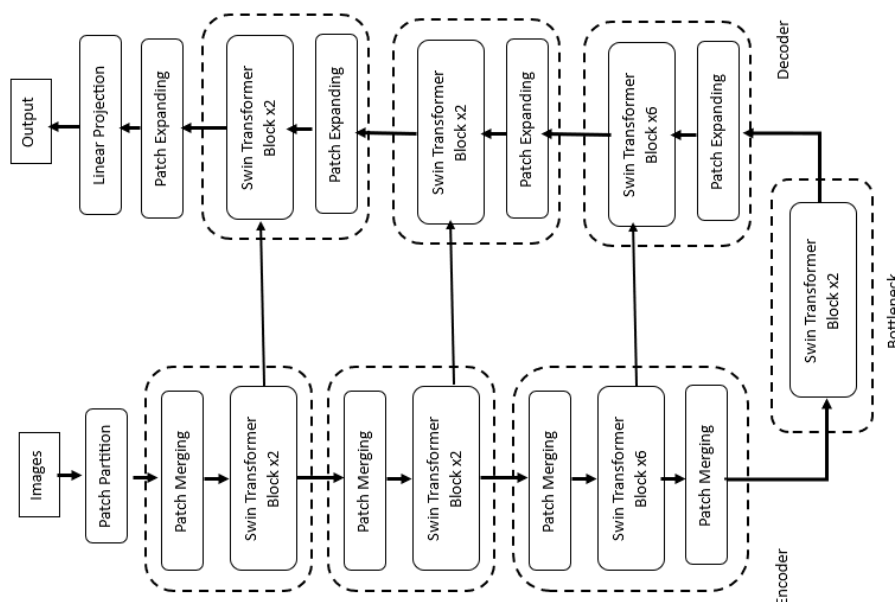
Bộ mã hóa trong kiến trúc Swin-Unet sử dụng ba giai đoạn đầu tiên của kiến trúc Swin Transformer phiên bản Swin-T, mỗi giai đoạn chứa các khối Swin Transformer con. Giai đoạn 1

và giai đoạn 2 mỗi giai đoạn chứa 2 khối Swin Transformer con, trong khi giai đoạn 3 chứa 6 khối. Các khối này được thiết kế để xử lý thông tin đặc trưng từ cấp độ thấp đến cấp độ cao, từ đó giảm dần kích thước tensor thông qua quá trình Patch Merging.

Nút cổ chai của mạng, nằm giữa bộ mã hóa và giải mã, sử dụng giai đoạn 4 của Swin Transformer phiên bản Swin-T, là giai đoạn có độ sâu lớn nhất, nhằm tối ưu hóa việc tổng hợp thông tin đặc trưng. Nơi đây, thông tin từ các lớp mã hóa được tổng hợp để tạo ra đặc trưng mức cao nhất trước khi bắt đầu quá trình giải mã.

Bộ giải mã của mô hình được xây dựng đối xứng với bộ mã hóa, sử dụng các khối Swin Transformer làm cốt lõi. Trong khi bộ mã hóa sử dụng lớp Patch Merging để giảm dần kích thước của bản đồ đặc trưng, bộ giải mã lại áp dụng lớp Patch Expanding để lấy mẫu nâng cấp các đặc trưng sâu. Lớp Patch Expanding này tái cấu trúc các bản đồ đặc trưng có kích thước liền kề, chuyển đổi chúng thành bản đồ đặc trưng có độ phân giải cao hơn. Skip connection từ bộ mã hóa được tích hợp thông qua lớp Concatenate, kết hợp thông tin đặc trưng từ các cấp độ khác nhau để hỗ trợ quá trình tái cấu trúc ảnh đầu ra. Điều này cho phép mô hình phục hồi chi tiết ảnh một cách chính xác, đồng thời duy trì thông tin quan trọng từ bộ mã hóa. Cuối cùng, lớp Dense được sử dụng để chiếu tensor đặc trưng về kích thước ban đầu của ảnh, và một lớp Activation với hàm kích hoạt 'sigmoid' được áp dụng để tạo ra bản đồ phân đoạn cuối cùng với kích thước $256 \times 256 \times 1$ như Hình 4.

Trong quá trình thiết kế mạng Swin-Unet, nghiên cứu đã áp dụng trọng số từ mô hình Swin Transformer V2 phiên bản 'tiny' (swin2_tiny_256) làm điểm khởi đầu. Mô hình này, đã được huấn luyện trước trên tập dữ liệu ImageNet, chứa đựng kiến thức về một loạt đặc trưng và mối quan hệ giữa các đối tượng trong lĩnh vực thị giác máy tính. Sự lựa chọn này nhằm mục đích tận dụng kiến thức đã học được từ một trong những tập dữ liệu lớn và đa dạng hiện nay, nhằm cải thiện khả năng của Swin-Unet trong việc trích xuất thông tin từ hình ảnh, đòi hỏi sự hiểu biết sâu sắc về cấu trúc và hình thái. Trọng số của 'swin2_tiny_256' được áp dụng như là điểm khởi đầu cho bộ mã hóa của SwinUnet. Điều này bao gồm việc khởi tạo các khối Swin Transformer với các trọng số đã học được từ ImageNet, giúp mô hình nhanh chóng thích ứng với dữ liệu mới và tiết kiệm đáng kể thời gian cần thiết cho việc huấn luyện mạng từ đầu.



Hình 4. Kiến trúc mô hình Swin-Unet được cài đặt dựa trên mô hình Swin-T

Trong nghiên cứu này sử dụng hàm tối ưu hóa Adam với tỉ lệ học là $5e-5$ và epsilon là $1e-5$. Kích thước lô (batch_size) đặt là 8. Ở đây mô hình tự dừng lại sau 150 epoch không có sự giảm về val_loss. Hàm mất mát sử dụng để huấn luyện ở đây là hàm tversky_loss.

2.4. Hàm mất mát Tversky loss

Trong nghiên cứu này, chúng tôi áp dụng hàm mất mát Tversky để phân đoạn hình ảnh MRI não, nhằm phát hiện vùng bất thường. Hàm mất mát Tversky là một biến thể của hàm mất mát Dice, được thiết kế để giải quyết vấn đề mất cân bằng lớp trong dữ liệu. Hàm này đặc biệt hữu ích trong việc phân đoạn các vùng bất thường, thường chiếm tỉ lệ nhỏ trong ảnh và dễ bị bỏ qua bởi các hàm mất mát thông thường.

Hàm mất mát Tversky được định nghĩa như công thức (1):

$$Tversky_{loss} = 1 - \frac{TP}{TP + \alpha FP + \beta FN} \quad (1)$$

Trong đó:

- TP (True Positives) là số lượng điểm ảnh được phân đoạn chính xác là vùng bất thường.
- FP (False Positives) là số lượng điểm ảnh bình thường bị nhầm lẫn là bất thường.
- FN (False Negatives) là số lượng điểm ảnh bất thường bị bỏ sót.

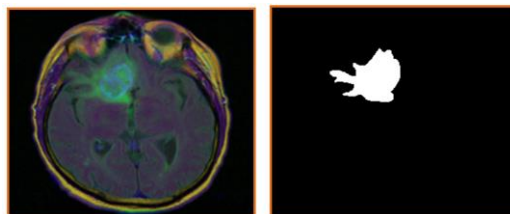
Các tham số α và β cho phép điều chỉnh độ nhạy của hàm mất mát đối với FP và FN. Trong nghiên cứu này, $\alpha = 0,3$ và $\beta = 0,7$ được áp dụng, điều này phản ánh mục tiêu ưu tiên giảm FN hơn là FP. Bởi vì, việc bỏ sót các vùng bất thường có thể dẫn đến hậu quả nghiêm trọng trong chẩn đoán và điều trị. Sự điều chỉnh này giúp tăng cường khả năng của mô hình trong việc phát hiện các vùng bất thường, ngay cả khi chúng chiếm tỉ lệ nhỏ trong ảnh.

3. Kết quả thực nghiệm

Trong bài báo này, các mô hình được xử lý trên một máy tính đơn với CPU Intel(R) Core(TM) i9-12900K 3,20 GHz đi kèm với 64GB ram cho phép huấn luyện với batch size lớn hơn, GPU có hiệu suất cao như GPU Nvidia GeForce RTX 4070 Ti giúp rút ngắn đáng kể thời gian tính toán. Mô hình Unet có số lượng tham số là: 8,635,809 và thời gian huấn luyện là 44s/epoch. Cải tiến mô hình Unet, ResUnet có số lượng tham số ít hơn (2,078,673 tham số) và thời gian huấn luyện nhanh hơn (30s/epoch). Mô hình SwinUnet có số lượng tham số cao nhất (36,071,623) và thời gian huấn luyện là 142s/epoch (gấp 4,73 lần mô hình ResUnet).

3.1. Tập dữ liệu

Tập dữ liệu LGG gồm 110 ảnh MRI não sử dụng chuỗi xung FLAIR cắt theo phương ngang với tổng 3929 lát cắt trong đó 1373 lát cắt có vùng bất thường và 2437 ảnh bình thường. Các lát cắt có chứa vùng bất thường sẽ có thêm file JSON chứa thông tin về vị trí các pixel chứa vùng bất thường gọi là mặt nạ vùng bất thường. Hình 5 minh họa ảnh gốc của lát cắt có chứa vùng bất thường và mặt nạ chứa vùng bất thường tương ứng.



Hình 5. Ảnh LGG và mặt nạ tương ứng

Tập dữ liệu Bệnh viện Trường Đại học Y Dược Cần Thơ (BV ĐHYDCT) gồm MRI não của 139 bệnh nhân sử dụng chuỗi xung FLAIR cắt theo phương ngang trung bình mỗi bệnh nhân có

20 lát cắt. Do bộ dữ liệu thu từ BV ĐHYDCT cũng như đa phần các bệnh viện ở các huyện xã tại Việt Nam đều sử dụng là ảnh MRI xám trong khi LGG là ảnh màu, vì vậy bước đầu tiên là chuyển ảnh LGG thành ảnh xám. Tỷ lệ ảnh bình thường trên tổng số ảnh của BV ĐHYDCT là khá lớn 3,7 trong khi LGG là 1,8. Vì vậy trong nghiên cứu này chúng tôi đã loại bỏ một số tập ảnh mà bệnh nhân không có lát cắt chứa vùng bất thường trong tập dữ liệu BV ĐHYDCT. Sau khi xử lý dữ liệu BV ĐHYDCT có 593 lát cắt bất thường và 1065 lát cắt bình thường [14].

3.2. Các thực nghiệm

3.2.1. Thực nghiệm 1: Tăng cường dữ liệu

Với tập dữ liệu thu thập từ BV ĐHYDCT, nhằm làm tăng độ chính xác của mô hình đề xuất, dữ liệu được tăng cường thông qua một loạt các thao tác: Ảnh được quay góc $\pm 10^\circ$, dịch chuyển ảnh theo chiều ngang, thay đổi kích thước 90-110%, lật ảnh, điều chỉnh độ sáng 10-110%, điều chỉnh độ tương phản 80-120%. Số lượng ảnh sau khi tăng cường và bổ sung tập LGG được chia ngẫu nhiên thành 3 phần: dữ liệu huấn luyện với 4466 ảnh, bộ dữ liệu điều chỉnh mô hình với 789 ảnh, và dữ liệu kiểm tra với 332 ảnh. Kết quả đánh giá trên 128 ảnh bất thường và 204 ảnh bình thường đạt được trình bày trong bảng 1, mô hình được tăng cường dữ liệu đã tăng về độ chính xác (Accuracy) từ 0,83 lên 0,88, chỉ số F1 có giá trị từ 0,74 tăng lên 0,82 so với mô hình không tăng cường dữ liệu. Điều này chứng tỏ việc tăng cường dữ liệu không chỉ giúp mô hình nhận diện chính xác vùng bất thường hơn mà còn giảm thiểu đáng kể số lượng dự đoán sai lệch đối với các ảnh chứa vùng bất thường.

Bảng 1. Bảng tổng hợp kết quả phân loại vùng bất thường

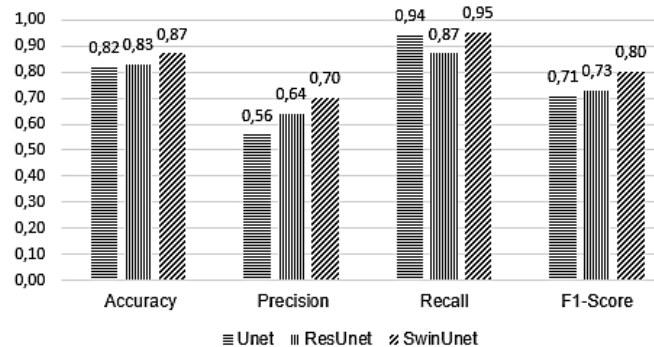
Phương pháp	TP (IoU>0)	FP	TN	FN	Accuracy	F1-score
Không tăng cường dữ liệu	81	47	193	11	0,83	0,74
Tăng cường dữ liệu	91	37	200	4	0,88	0,82

3.2.2. Thực nghiệm 2: Đánh giá phân chia dữ liệu theo ảnh (lát cắt)

Thực nghiệm đánh giá mô hình Unet, ResUnet và Swin-Unet trên bộ dữ liệu được phân chia theo ảnh có kết quả trình bày trong Bảng 2, Bảng 3 và Hình 6. Trong bộ dữ liệu của ĐHYDCT lấy ra 20% ảnh làm bộ kiểm tra, trong 80% còn lại sẽ gộp với bộ dữ liệu LGG và tiếp tục lấy ra 15% dữ liệu ảnh vừa gộp để làm bộ xác thực và số còn lại làm bộ huấn luyện. Dữ liệu ảnh sẽ được lấy một cách ngẫu nhiên ba lần với các trường hợp random_state lần lượt là 50, 250 và 550.

Bảng 2. Tổng hợp kết quả phát hiện và phân loại vùng bất thường

Random State	Mô hình	Accuracy	Precision	F1-score	TP (IoU>0)	FP	TN	FN
50	Unet	0,83	0,58	0,72	72	52	203	5
	ResUnet	0,82	0,63	0,72	78	46	194	14
	SwinUnet	0,87	0,69	0,79	85	39	203	5
250	Unet	0,83	0,59	0,73	76	52	201	3
	ResUnet	0,85	0,70	0,78	90	38	192	12
	SwinUnet	0,86	0,70	0,80	89	39	198	6
550	Unet	0,80	0,52	0,67	67	61	198	6
	ResUnet	0,81	0,58	0,70	74	54	195	9
	SwinUnet	0,88	0,71	0,82	91	37	200	4



Hình 6. Biểu đồ so sánh giá trị trung bình các lần huấn luyện

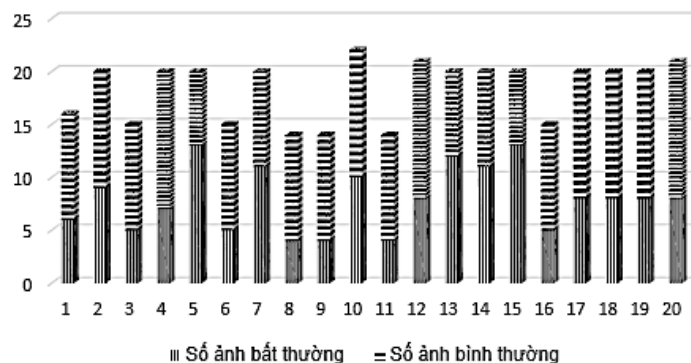
Thực nghiệm 2 cho thấy Swin-Unet cho kết quả tốt hơn Unet và ResUnet với Accuracy cao nhất là 0,88 và Precision 0,71 trong phân loại, cùng với mIoU và mDSC lần lượt là 0,91 và 0,94 trong phân đoạn. Kết quả cho thấy, Swin-Unet hiệu quả trong cả hai tác vụ phân loại và phân đoạn, với sự cân bằng tối ưu giữa việc giảm dương tính giả và khả năng xác định chính xác vùng bất thường. ResUnet cũng thể hiện khả năng cải thiện Precision trong phân loại và duy trì hiệu suất phân đoạn tốt, trong khi Unet cần tập trung vào việc giảm số lượng dương tính giả trong phân loại để tăng cường hiệu suất tổng thể.

Bảng 3. Kết quả đánh giá phân đoạn vùng bất thường

Random State	Mô hình	mIoU	mDSC	mPrecision	mRecall
50	Unet	0,90	0,93	0,97	0,91
	ResUnet	0,88	0,91	0,96	0,90
	SwinUnet	0,86	0,91	0,97	0,88
250	Unet	0,90	0,93	0,98	0,91
	ResUnet	0,87	0,91	0,96	0,89
	SwinUnet	0,88	0,92	0,96	0,91
550	Unet	0,91	0,94	0,98	0,92
	ResUnet	0,90	0,93	0,97	0,92
	SwinUnet	0,91	0,94	0,98	0,92

3.2.3. Thực nghiệm 3: Đánh giá phân chia dữ liệu theo bệnh nhân

Thực nghiệm này chỉ lấy được 106/139 ảnh MRI (do 33 ảnh MRI không có vùng bất thường) trong dữ liệu BV ĐHYDCT, trong đó có 20 ảnh làm kiểm tra như hình 7 (159 ảnh bất thường và 208 ảnh bình thường), 86 bộ ảnh còn lại gộp với dữ liệu LGG để huấn luyện (1606 ảnh bất thường và 3013 ảnh bình thường) và xác thực (201 ảnh bất thường và 208 ảnh bình thường).



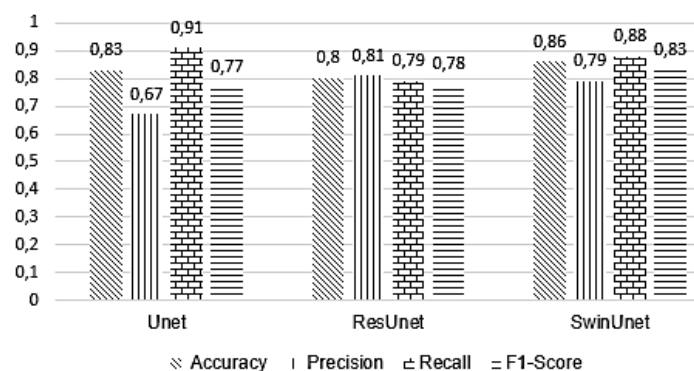
Hình 7. Số lượng ảnh bất thường và bình thường trên 20 ảnh kiểm tra

Hiệu quả của mô hình được trình bày trong bảng 4, bảng 5 và hình 8. Kết quả cho thấy Unet có sự cân bằng giữa việc phát hiện và xác định vị trí vùng bất thường. ResUnet có Accuracy thấp hơn (0,8) nhưng Recision cao (0,81) cho thấy giảm dương tính giả tốt hơn. Sự chênh lệch giữa hiệu suất phân loại và phân đoạn cho thấy ResUnet có thể cần cải thiện khả năng xác định vị trí và diện tích chính xác của vùng bất thường. Swin-Unet thể hiện hiệu suất ổn định và cao nhất trong cả phân loại với Accuracy 0,86, Precision 0,79 và F1-score 0,83, và phân đoạn với mIoU 0,84 và mDSC 0,89. Swin-Unet hiệu quả trong phát hiện vùng bất thường mà còn xác định chính xác về vị trí và diện tích của chúng.

Thực nghiệm 3 cho thấy Swin-Unet có sự cân bằng tốt nhất giữa hiệu suất phân loại và phân đoạn, làm nổi bật khả năng tổng quát hóa và độ chính xác cao trong cả hai tác vụ. Unet, mặc dù có hiệu suất phân loại tốt nhưng trong phân đoạn vẫn có chỗ cần cải thiện so với Swin-Unet. ResUnet cho thấy Precision cao trong phân loại nhưng lại không duy trì được hiệu suất phân đoạn hiệu quả như hai mô hình còn lại. Kết quả này nhấn mạnh tầm quan trọng của việc lựa chọn mô hình phù hợp dựa trên nhu cầu cụ thể của nghiên cứu và ứng dụng, đặc biệt khi xem xét tới việc cân bằng giữa khả năng phát hiện và xác định chính xác các vùng bất thường trên ảnh MRI não.

Bảng 4. Bảng tổng hợp kết quả phân loại vùng bất thường

Mô hình	Accuracy	Precision	F1-Score	TP (IoU>0)	FP	TN	FN
Unet	0,83	0,67	0,77	106	53	197	11
ResUnet	0,80	0,81	0,78	128	31	165	43
SwinUnet	0,86	0,79	0,83	126	33	191	17



Hình 8. Biểu đồ so sánh kết quả phân loại

Bảng 5. Kết quả đánh giá phân đoạn vùng bất thường

Mô hình	mIoU	mDSC	mPrecision	mRecall
Unet	0,86	0,90	0,96	0,88
ResUnet	0,81	0,86	0,88	0,89
Swin-Unet	0,84	0,89	0,95	0,87

4. Kết luận

Việc phát hiện kịp thời khối u hỗ trợ các bác sĩ trong quá trình chẩn đoán và điều trị cho bệnh nhân được thực hiện hiệu quả trong tình trạng các bệnh viện luôn quá tải là rất cần thiết. Trong nghiên cứu này, mô hình Swin-Unet được thực nghiệm và so sánh đánh giá với mô hình Unet và ResNet. Mô hình đề xuất được huấn luyện trên tập dữ liệu LGG của 110 bệnh nhân với 3929 ảnh MRI (2437 ảnh không chứa vùng bất thường, 1373 ảnh có chứa vùng bất thường) kết hợp với tập dữ liệu đã được thu thập và gán nhãn của 106 bệnh nhân (1551 ảnh bình thường và 604 ảnh bất thường) tại Bệnh viện Trường Đại học Y Dược Cần Thơ. Kết quả cho thấy mô hình Swin-Unet với một số tùy chỉnh khi áp dụng vào ảnh MRI não cũng đạt hiệu quả cao hơn nghiên cứu trước

đó [12] cho ảnh CT bụng và MRI tim. Qua đó nghiên cứu đã làm rõ được các điểm mạnh cũng như hạn chế của từng mô hình và cung cấp một cái nhìn tổng quát về cách thức mà các mô hình này có thể được tối ưu hóa để được áp dụng trong thực tế vấn đề phát hiện vùng bất thường đối với ảnh y khoa. Đặc biệt, mô hình Swin-Unet hiệu quả trong việc xác định vùng bất thường trên ảnh MRI não nhờ có độ chính xác cao và ổn định trong phân loại ảnh có bất thường hay không và sau đó là phân đoạn vùng tổn thương nhưng thời gian huấn luyện cao hơn mô hình Unet và ResUnet lần lượt 3,7 và 4,3 lần. Phương pháp đánh giá theo bệnh nhân hay theo ảnh cũng làm thay đổi kết quả đánh giá hiệu quả của giải thuật.

Việc thu thập thêm nhiều bộ ảnh y tế ở nhiều cơ sở y tế khác để làm dữ liệu huấn luyện trong tương lai kết hợp tối ưu hóa kỹ thuật và tham số trong mô hình để có thể áp dụng chẩn đoán vùng tổn thương trong thực tế cho các bệnh viện chưa có điều kiện có những thiết bị tiên tiến.

Lời cảm ơn

Chúng tôi xin trân trọng cảm ơn đến nhóm Bác sĩ khoa Chẩn đoán hình ảnh Bệnh viện Trường Đại học Y Dược Cần Thơ đã hỗ trợ gán nhãn hoàn thiện bộ dữ liệu.

TÀI LIỆU THAM KHẢO/ REFERENCES

- [1] Tsai, A. Yezzi, W. Wells, C. Tempany, D. Tucker, A. Fan, W. Grimson, and A. Willsky, "A shape-based approach to the segmentation of medical imagery using level sets," *IEEE Transactions on Medical Imaging*, vol. 22, no. 2, pp. 137–154, 2003.
- [2] K. Held, E. Kops, B. Krause, W. Wells, R. Kikinis, and H.-W. Muller-Gartner, "Markov random field segmentation of brain mr images," *IEEE Transactions on Medical Imaging*, vol. 16, no. 6, pp. 878–886, 1997.
- [3] M. L. Le, N. M. T. Tran, T. N. Ho, and N. K. Pham, "Using U-Net model to detect abnormal areas on brain MRI images," *The National Conference on Fundamental and Applied IT Research (in Vietnamese) (FAIR)*, Nha Trang, 2020, pp. 656–662.
- [4] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, ser. LNCS, vol. 9351, Springer, 2015, pp. 234–241.
- [5] X. Xiao, S. Lian, Z. Luo, and S. Li, "Weighted res-unet for high-quality retina vessel segmentation," *2018 9th International Conference on Information Technology in Medicine and Education (ITME)*, 2018, pp. 327–331.
- [6] Z. Zhou, M. R. Siddiquee, N. Tajbakhsh, and J. Liang, *Unet++: A nested u-net architecture for medical image segmentation*. Springer Verlag, 2018, pp. 3–11.
- [7] X. Li, H. Chen, X. Qi, Q. Dou, C.-W. Fu, and P.-A. Heng, "H-denseunet: Hybrid densely connected unet for liver and tumor segmentation from ct volumes," *IEEE Transactions on Medical Imaging*, vol. 37, no. 12, pp. 2663–2674, 2018.
- [8] O. Çiçek, A. Abdulkadir, S. Lienkamp, T. Brox, and O. Ronneberger, "3d u-net: Learning dense volumetric segmentation from sparse annotation," in *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, ser. LNCS, vol. 9901, Springer, Oct. 2016, pp. 424–432.
- [9] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," *CoRR*, 2020, doi: 10.48550/arXiv.2005.12872.
- [10] Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," in *International Conference on Learning Representations*, arXiv preprint arXiv:2010.11929, 2020.
- [11] A. Hatamizadeh, V. Nath, Y. Tang, D. Yang, H. R. Roth, and D. Xu, "Swin UnetR: Swin Transformers for Semantic Segmentation of Brain Tumors in MRI Images," in *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries*, Springer International Publishing, 2022, pp. 272–284.
- [12] H. Cao, Y. Wang, J. Chen, D. S. Jiang, X. P. Zhang, Q. Tian, and M. Wang, "Swin-Unet: Unet-like Pure Transformer for Medical Image Segmentation," in *European conference on computer vision*. Cham: Springer Nature Switzerland, 2022, pp. 205–218.
- [13] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swintransformer: Hierarchical vision transformer using shifted windows," in *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, Montreal, QC, Canada, 2021, pp. 9992–10002.
- [14] N. M. T. Tran, T. L. Bui, M. L. Le, T. N. Phu, and N. K. Pham, "Applying CGAN model for detecting abnormal areas on brain MRI images," *The National Conference on Fundamental and Applied IT Research (in Vietnamese) (FAIR)*, Ha Noi, 2022, pp. 462–469.