Brain Tumour Segmentation Using Deep Neural Networks

Aishik Biswas
Student Id: 1143769
abiswas3@lakeheadu.ca
Lakehead University

Yenna Dinesh
Student Id: 1145701
dyenna@lakeheadu.ca
Lakehead University

Abstract

The best medical picture segmentation techniques in recent years have been UNet and its most recent extensions, such as TransUNet. These networks are parameter-heavy, computationally difficult, and slow to utilise, hence they cannot be efficiently used for quick image segmentation in point-of-care applications. So we intended to develop two new and distinct models independently, train them on the same dataset, and then create a combined model to test their accuracy. We chose a complicated model, the Swinunetunet model, which has a higher computational complexity, inference time, and number of parameters. On the other hand, we used the Unext model to solve this challenge and create an efficient network with reduced computational cost, fewer parameters, and a faster inference time while still keeping high performance. So we combined these two models to form a merged model and compared the results.

1. Introduction

Image segmentation is a critical issue in image processing and computer vision, with several applications including scene interpretation, medical picture analysis, robotic perception, video surveillance, augmented reality, and image compression. Several picture segmentation techniques have been developed in the literature. Due to the success of deep learning models in a variety of vision applications, there has recently been a significant amount of effort directed towards creating picture segmentation algorithms utilising deep learning models.

Brain tumours are among the most dangerous forms of malignancies in the world. Glioma, the most common primary brain tumour, develops as a result of glial cell carcinogenesis in the spinal cord and brain. Glioma has many histological and malignancy grades, with glioblastoma patients having an average survival duration of less than 14 months following diagnosis. Magnetic Resonance Imaging (MRI), a common non-invasive method, generates a vast and diverse number of tissue contrasts in each imaging modality and is frequently utilised by medical professionals to identify brain cancers. However, manual segmentation and interpretation of structural MRI images of brain tumours is a difficult and time-consuming process that has so far been limited to experienced neuroradiologists.

Currently, despite the fact that CNN-based algorithms have achieved good performance in the field of medical picture segmentation, they cannot fully fulfil the stringent segmentation accuracy criteria of medical applications. Image segmentation remains a difficult task in medical image analysis. Because of the intrinsic locality of convolution operations, CNN-based techniques struggle to learn explicit global and long-range semantic information exchange. Motivated by the success of the Swin Transformer, we sought to use Swin-Unet to utilise the capability of Transformer for 2D medical picture segmentation in this study.

Many transformer-based networks have recently been proposed for medical image segmentation because they develop a global understanding of pictures that can aid with segmentation. The TransUNet design transforms the ViT architecture [into a UNet for 2D medical picture segmentation]. Other transformer-based networks for medical picture segmentation have also been presented. It should be noted that practically all of the preceding efforts have concentrated on improving network performance while paying little attention to computational complexity, inference time, or the number of parameters, all of which are important in many real-world applications. Because the majority of them are used for laboratory analysis, they are evaluated on machines with significant computing capability (like GPUs).
Medical imaging solutions have recently been translated from the laboratory to the bedside. Because the testing and analysis are performed at the patient's side, this is known as point-of-care imaging. Point-of-care imaging [23] provides doctors with more service alternatives and better pa-

tient care. It aids in minimising the time and processes required for patients to attend radiology institutions. In this study, we focus on tackling this challenge and designing an efficient network with reduced computational cost, fewer parameters, a faster inference time, and high performance. Designing such a network is critical in order to meet the changing trends in medical imaging from laboratory to bedside.

To that purpose, we sought to implement UNeXt, which is built with convolutional networks and MLPs (multilayer perceptron).

2. Literature Review

- **A Review on Convolutional Neural Networks for Brain Tumor Segmentation: Methods, Datasets, Libraries, and Future Directions,M.K.Balwant Dept. of Computer Science, School of Science, UP Rajarshi Tandon Open University Prayagraj, Uttar Pradesh, India**
  This paper investigated the automated segmentation of brain tumours from MRI images using current advances in CNN-based techniques, which is based on data from multiple previous publications. It investigates common deep learning (DL) frameworks and tools for the quick and easy development of CNN models. The development region is also outlined, and the present DL designs are appraised critically. More than 50 scholarly publications from 2014 to 2020 were retrieved using Google Scholar and PubMed for this approach.

  Also retrieved are the key publications relating to their work, as well as proceedings from significant conferences such as MICCAI, MIUA, and ECCV. This study looked into many yearly challenges connected to this topic, such as the Multimodal Brain Tumor Segmentation Challenge (MICCAI BRATS) and the Ischemic Stroke Lesion Segmentation Challenge (ISLES). After doing a thorough literature search on the subject, they discovered that there are primarily three types of CNN architecture for brain tumour segmentation: single-path and multi-path, fully convolutional, and cascaded CNNs.

- **Fully Convolutional Networks for Semantic Segmentation-Jonathan Long, Evan Shelhamer, Trevor Darrell**
  They described how convolutional networks may be used to handle exceedingly complicated prediction problems, how to construct and characterise the space of totally convolutional networks, and how to link to earlier models. In order to transfer their learnt representations to the segmentation problem, they convert contemporary classification networks (AlexNet, VGG net, and GoogLeNet) into fully convolutional networks. They then build a skip architecture that produces precise and comprehensive segmentations.

  They also demonstrated that convolutional networks trained end-to-end, pixel-by-pixel, outperform the state-of-the-art in semantic segmentation. Their major breakthrough is where we can design ”completely convolutional” networks that can accept input of any size and produce output of any size with efficient inference and learning.

- **U-Net: Convolutional Networks for Biomedical Image Segmentation-Olaf Ronneberger, Philipp Fischer, Thomas Brox**
  They build on a more elegant design, the so-called ”fully convolutional network,” in this research. They improved and extended this architecture so that it can function with less training photos and produce more precise segmentations; they also provided examples of how the u-net may be utilised for different segmentation tasks.

  They also demonstrated how the training strategy and unet network rely largely on data augmentation to make the best use of the annotated examples provided. They proved that such a network can be trained from scratch using only a few pictures, beating the previous top approach. The design is made up of a contracting path for capturing context and a symmetric expanding path for exact localization.

- **TransUNet: Transformers Make Strong Encoders for Medical Image Segmentation-Jieneng Chen, Yongyi Lu, Qihang Yu, Xiangde Luo, Ehsan Adeli, Yan Wang, Le Lu, Alan L. Yuille, Yuyin Zhou**
  In this study, they suggested TransUNet as a robust alternative for medical image segmentation, which benefits from both Transformers and U-Net. On the one hand, as the input sequence for extracting global contexts, the Transformer encodes tokenized image patches from a CNN feature map. The decoder, on the other hand, upsamples the encoded features, which are then coupled with the high-resolution CNN feature maps to allow for exact localisation.

  They contend that Transformers may be used as powerful encoders for medical image segmentation tasks, with the addition of U-Net to improve finer details by retrieving localised spatial information. TransUNet

outperforms alternative approaches in a variety of medical applications, including multi-organ segmentation and cardiac segmentation.

- **Cross-Modal Self-Attention Network for Referring Image Segmentation-Linwei Ye, Mrigank Rochan, Zhi Liu, Yang Wang**
  They introduced a cross-modal self-attention (CMSA) module in this study that successfully captures the long-term connections between verbal and visual characteristics, resulting in an improved feature representation to focus on crucial information for referenced entities. Their model also focus on informative words in the referring expression and relevant regions in the input image in an adaptive manner.

  They also presented a gated multi-level fusion module for selectively integrating self-attentive cross-modal information corresponding to distinct picture levels. This module manages the information flow of features at various levels. They tested the suggested method on four different datasets. Furthermore, the proposed gated multi-level fusion module fuses characteristics from multiple levels adaptively via learnable gates for each individual level.

3. Proposed method:

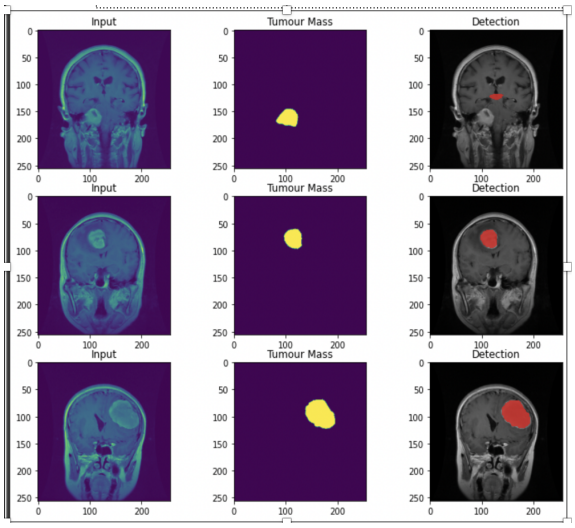
UNet is the foundation model used by all researchers in the field of medical picture segmentation. Although Unet is the foundation model for medical image segmentation, several advancements and models have emerged in recent years, including TransUnet, TransBTS, SegNet, attention UNet, UNet++, UNet 3+, V-Net, and many more. Every model is constructed differently, although transformers and attention mechanisms are the key components of models these days. Image segmentation is an encoder-decoder architecture, and the models developed these days include transformer-based models, attention-based models, or a mix of the two.

The technique we developed is a merged model that combines two extremely popular and real-world-based models, SwinUNet and UNeXt. TransBTS, TransUNet, and other models offer high accuracy and good dice scores, but the drawback is that they are computationally costly, which takes a long time and cannot be perfect for real-world applications that require rapid and exact outputs. When two models are combined to produce one model, a merged or ensemble model is the ideal choice for real-world applications since it has a short training time, a faster inference time, and greater accuracy. As a result, we propose an ensembel model of SwinUNet and UNeXt.

4. Dataset:

This brain tumour dataset contains 3064 T1-weighted contrast-enhanced images from 233 patients with three different types of brain tumours: meningioma (708 slices), glioma (1426 slices), and pituitary tumour (1426 slices) (930 slices). Due to the repository's file size constraint, we divided the dataset into four subsets and packaged it as four.zip files, each comprising 766 slices. The indices of 5-fold cross-validation are also supplied. This information has been arranged in matlab data format (.mat file). Each file stores a struct containing the following fields for an image:

- cjdata.label: 1 for meningioma, 2 for glioma, 3 for pituitary tumor
- cjdata.PID: patient ID
- cjdata.image: image data
- cjdata.tumorBorder: a vector storing the coordinates of discrete points on tumor border. For example, [x1, y1, x2, y2,...] in which x1, y1 are planar coordinates on tumor border.It was generated by manually delineating the tumor border. So we can use it to generate binary image of tumor mask.
- cjdata.tumorMask: a binary image with 1s indicating tumor region



5. Swin-Unet Model:

Swin-Unet consists of encoders, decoders, and skip connections. Swin Transformer blocks are the fundamental unit of Swin-Unet. The medical images are separated into non-overlapping patches with a patch size of 4*4 for the encoder

to turn the inputs into sequence embeddings. By using this partitioning method, the feature dimension of each patch is reduced to 4*4*3 = 48. In addition, a linear embedding layer is used to convert the projected feature dimension into an arbitrary dimension (represented as C). To build the hierarchical feature representations, the changed patch tokens are passed via numerous Swin Transformer blocks and patch merging layers.

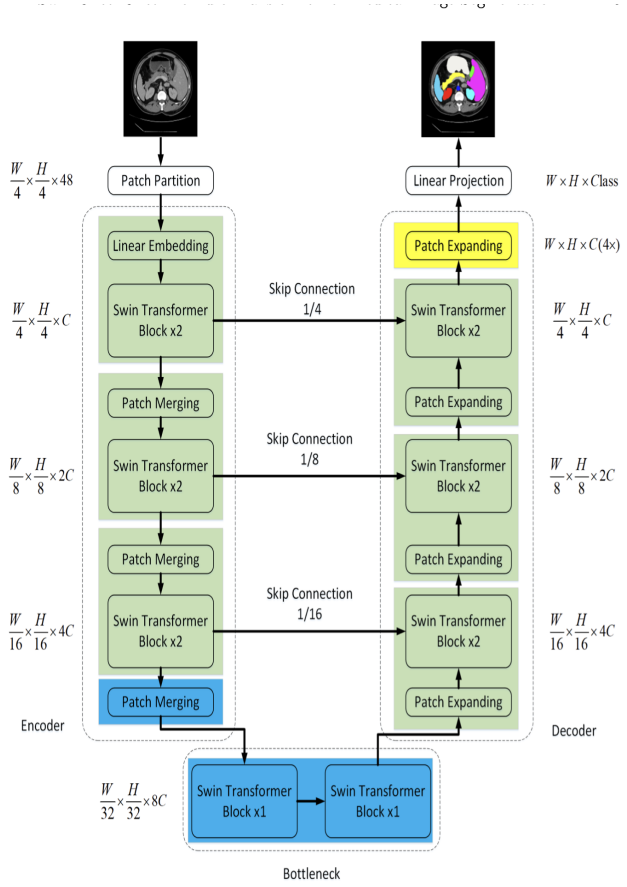


Figure 1. Swin-Unet Architecture

The patch merging layer manages with downsampling and increasing dimension, whereas the Swin Transformer block is in responsible for feature representation learning. We create a symmetric transformer-based decoder inspired by U-Net. Swin Transformer block and patch expanding layer make up the decoder. To compensate for the loss of spatial information induced by downsampling, the recovered context features are fused with multiscale features from the encoder through skip connections. A patch expanding layer, as opposed to a patch merging layer, is specifically designed to do up-sampling. The patch expanding layer reshapes nearby dimension feature maps into big feature maps with two upsamplings of resolution.Finally, the final patch expansion layer is utilised to conduct 4 up-sampling to restore the resolution of the feature maps to the input resolution (W*H), and a linear projection layer is used to these up-sampled features to generate the pixel-level segmentation predictions.

5.1. Swin Transformer block:

Unlike the traditional multi-head self-attention (MSA) module, the swin transformer block is built on shifted windows. Two swin transformer blocks are shown in succession. Each swin transformer block is made up of a LayerNorm (LN) layer, a multi-head self-awareness module, a residual connection, and a 2-layer MLP with GELU non-linearity. In the two succeeding transformer blocks, the window-based multi-head self attention (W-MSA) module and the shifted window-based multi-head self attention (SW-MSA) module are used.

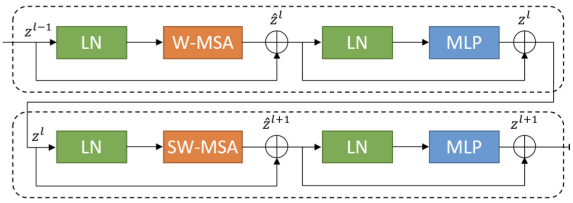


Figure 2. Swin Transformer Block

Based on such window partitioning mechanism, continuous swin transformer blocks can be formulated as:

$$\hat{z}^{l} = W\text{-}MSA(LN(z^{l-1})) + z^{l-1},$$

$$z^{l} = MLP(LN(\hat{z}^{l})) + \hat{z}^{l},$$

$$\hat{z}^{l+1} = SW\text{-}MSA(LN(z^{l})) + z^{l},$$

$$z^{l+1} = MLP(LN(\hat{z}^{l+1})) + \hat{z}^{l+1},$$

$$Attention(Q, K, V) = SoftMax(\frac{QK^{T}}{\sqrt{d}} + B)V,$$

Figure 3. where Q,K,V RM̂×d denote the query, key and value matrices. M2 and d represent the number of patches in a window and the dimension of the query or key, respectively. And, the values in B are taken from the bias matrix Bˆ R(2M 1)×(2M +1)

5.2. Encoder:

The C-dimensional tokenized inputs with the resolution of H4 W4 are fed through two successive Swin Transformer blocks in the encoder to conduct representation learning, while the feature dimension and resolution remain unchanged. Meanwhile, the patch merging layer will reduce the number of tokens (2 *downsampling) and raise the feature dimension to double its original size. In the encoder, this technique will be performed three times.

5.3. Patch merging layer:

The patch merging layer divides the input patches into four sections and connects them together. The feature resolution will be downsampled by 2 as a result of this procedure. Furthermore, because the concatenate operation causes the feature dimension to increase by 4, a linear layer is applied to the concatenated features to unify the feature dimension to the initial dimension of 2.

5.4. Decoder:

The symmetric decoder, like the encoder, is implemented using the Swin Transformer block. To that purpose, we employ the patch expanding layer in the decoder rather than the patch merging layer in the encoder to up-sample the retrieved deep features. The patch expanding layer reshapes neighbouring dimension feature maps into a higher resolution feature map (2 *up-sampling) and decreases the feature dimension to half of the original dimension.

5.5. Patch expanding layer:

As an example, before upsampling, a linear layer is applied to the input features (W/32 *H/32 *8C) to increase the feature dimension to twice the original dimension (W/32 *H/32*16C). Then, we utilise the rearrange operation to double the resolution of the input features and lower the feature dimension to a quarter of the input dimension (W/32 *H/32 *16C W/16 *H/16*4C).

5.6. Skip Connection:

The skip connections, like the U-Net, are used to fuse the encoder's multi-scale features with the up-sampled features. To limit the loss of spatial information induced by down sampling, we concatenate the shallow and deep features together. The dimension of the concatenated features is kept the same as the size of the up- sampled data after a linear layer.

We have achieved training and validation loss for 10 Epochs coming to:

- Training loss: 0.0085
- Validation loss: 0.0490
- Training dice_coef: 0.8323
- Validation dice_coef: 0.5324

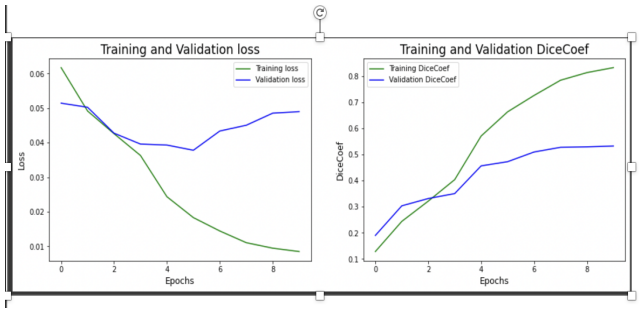


Figure 4. Evaluation results of Swin-Unet

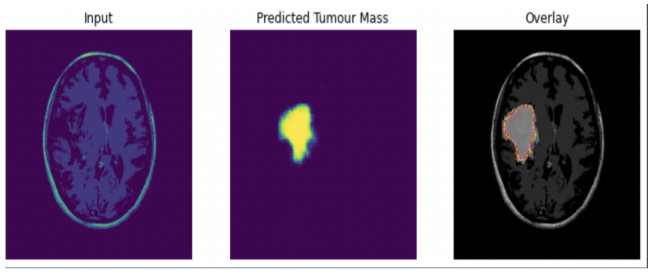


Figure 5. Predicted output

6. Unext Model:

UNeXt is an encoder-decoder architecture comprised of two stages: 1) Convolutional and 2) Tokenized MLP. The input picture is transmitted via the encoder, which has three convolutional blocks and two Tokenized MLP blocks. The decoder consists of two Tokenized MLP blocks followed by three convolutional blocks. Each encoder block decreases feature resolution by two, whereas each decoder block raises feature resolution by two. There are also skip connections between the encoder and decoder.

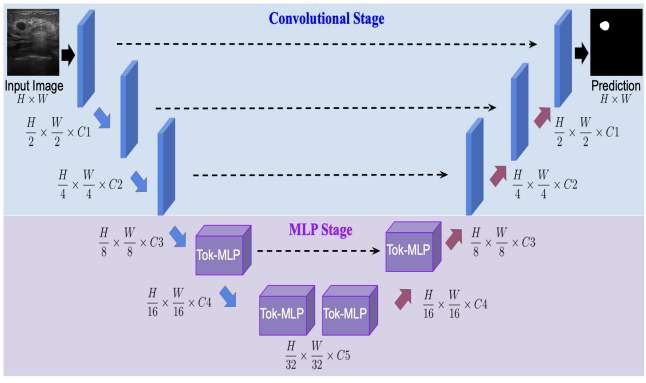


Figure 6. Unext Architecture

6.1. Convolutional Block:

Each convolutional block has a convolution layer, a batch normalisation layer, and ReLU activation. We utilise a kernel size of three, a stride of one, and padding of one. To upsample the feature maps, the encoder conv blocks employ a max-pooling layer with pool window 2 * 2, whereas the decoder conv blocks use a transpose convolution layer. Instead of bilinear interpolation, we chose transpose convolution since it is essentially learnable upsampling and adds to additional learnable parameters.

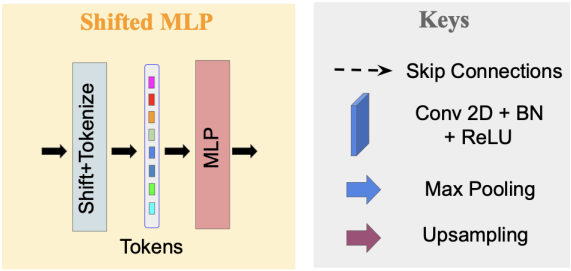


Figure 7.

6.2. Shifted MLP:

Before tokenizing, we move the axis of the channels of conv features in shifted MLP. This allows the MLP to focus on only a subset of the conv features, inducing locality to the block. The notion is similar to that of the Swin transformer, in which window-based attention is used to provide additional locality to an otherwise totally global model. Because the Tokenized MLP block contains two MLPs, the features are shifted across width in one and across height in the other, as in axial-attention. We divide the features into h divisions and move them by j places along the chosen axis. This allows us to construct random windows with locality along an axis.

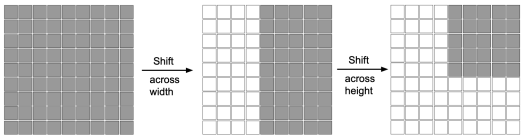
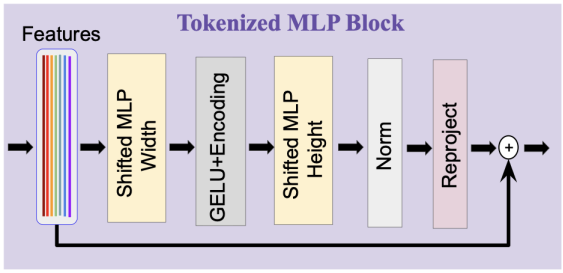


Figure 8. Shifting operations. To induce window locality in the network, the features are moved successively across height and width before tokenizing.

6.3. Tokenized MLP Stage:

We first shift the features and project them into tokens in the tokenized MLP block. To tokenize, we begin with a kernel size of 3 and increase the number of channels to E, where E is the embedding dimension (number of tokens), which is a hyperparameter. The tokens are subsequently sent to a shifted MLP (across width), whose hidden dimensions are a hyperparameter H. Following that, the features are processed via a depth-wise convolutional layer (DW- Conv).

DWConv employs fewer parameters, which boosts efficiency. Then we apply a GELU activation layer. We utilise GELU instead of RELU since it is a smoother option that performs better. Furthermore, contemporary designs like as ViT and BERT have effectively leveraged GELU to get better outcomes. The features are then sent through a second shifted MLP (across height) that changes the dimensions from H to O. In this case, we employ a residual connection and add the original tokens as residuals. The output features are then sent to the next block after we apply layer normalisation (LN). LN is chosen over BN because it makes more sense to normalise along the tokens rather than across the batch in the Tokenized MLP block.



We have achieved training and validation loss for 10 Epochs coming to:

- Training loss: 0.0109
- Validation loss: 0.0429
- Training dice_coef:0.7883
- Validation dice_coef: 0.5184

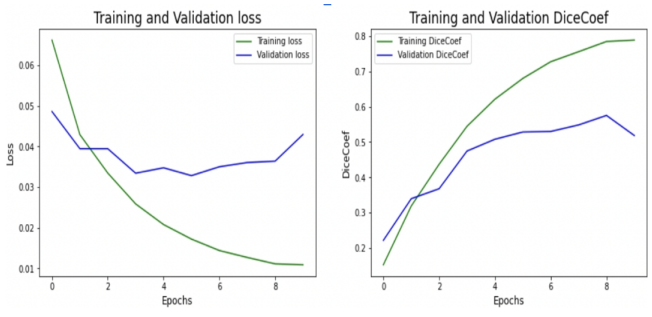


Figure 9. Evaluation results of Unext

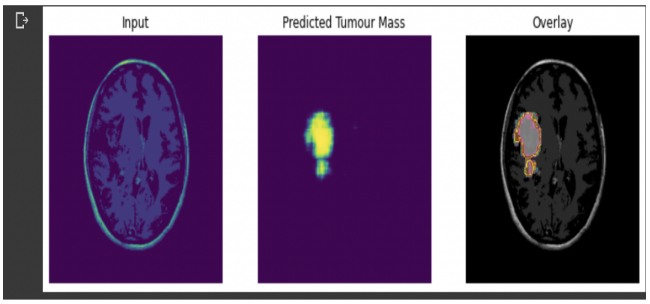


Figure 10. Predicted output

7. Merged Model:

Ensemble methods are ways for developing several models and then combining them to obtain better results. Ensemble approaches often yield more accurate results than a single model. Ensemble learning may also be used to develop a new model by combining the functionality of many deep learning models. Creating a new model offers numerous advantages over training a new model from start. Because most learning is generated from combined models, training the combined model requires relatively little data.

- Constructing a merged model requires less time than building a fresh model.
- When models are integrated, computer resources are reduced.
- New integrated models offer more accuracy and capabilities than those used to create the new model

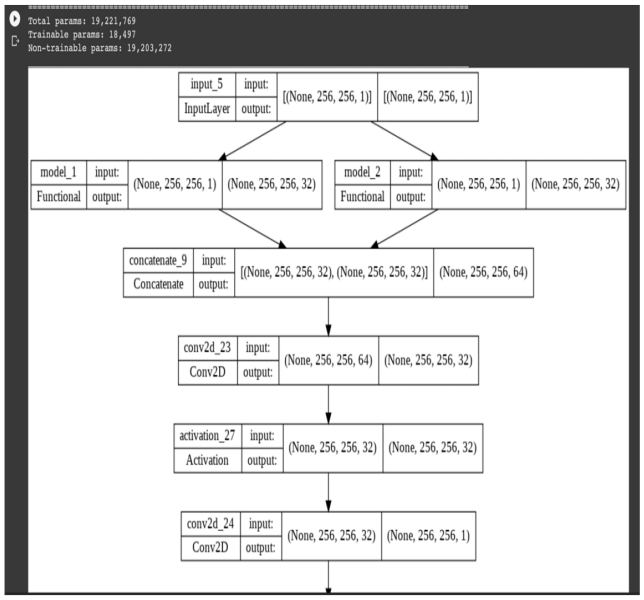


Figure 11. Architecture of the concatenated model

Concatenate was used to join layers of Swin-Unet and Unext modes. 'concatenate' is a built-in technique in Tensorflow packages (tf.keras.layers). We combined two models' layers and added two more CNN layers to achieve improved accuracy and lower computational cost in order to share information and learn new features based on both models that are leading to the segmentation job. We made certain that both models had the same input dimensions since 'concatenation' demands that all dimensions save the concatenation axis be of the same form.

We have achieved training and validation loss for 10 Epochs coming to:

- Training loss: 0.0069
- Validation loss: 0.0313
- Training dice_coef: 0.8859
- Validation dice_coef: 0.6799

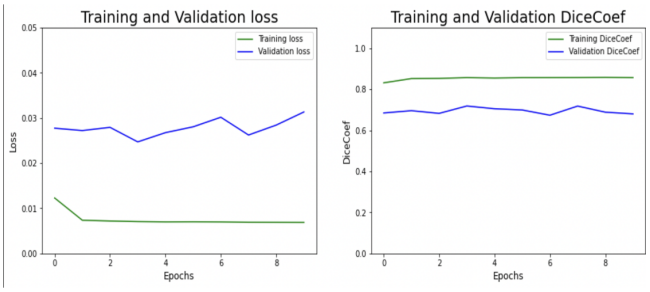


Figure 12. Evaluation results of Merged Model

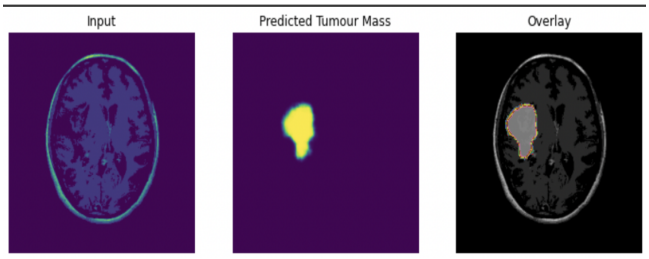


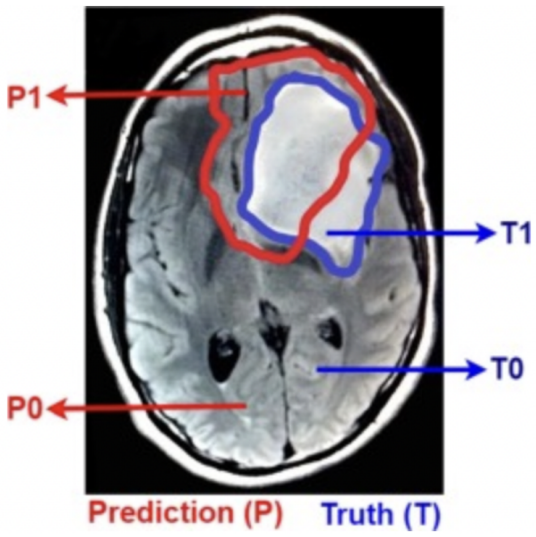
Figure 13. Predicted output

8. Evaluation Metrics:

For measuring the segmentation performance of automated techniques, four metrics are widely used: Dice score, sensitivity (true positive rate), specificity (true negative rate), and the Hausdorff distance. We used Dice score (DSC) which is a measure of the overlap between the ground truth and forecasted areas in relation to the total area

of both. A Dice score of 1 shows perfect overlap or segmentation, whereas a Dice score of 0 indicates no overlap or poor segmentation.

Dice score is determined as two times the area of overlap divided by the total number of pixels in both sections using the equation below. We receive T0, T1 as annotated by medical professionals and P0, P1 as predicted by an automated technique for each of the three tumour locations.



$$Dice(P,T) = \frac{|P_1 \cap T_1|}{(|P_1| + |T_1|)/2}$$

Figure 14. After segmentation, the distinct areas of the brain are labelled T0, T1, P0, and P1. Dice score, sensitivity, specificity, and robust Hausdorff distance are all calculated using these areas. The blue outline zone identified with T1 represents the genuine tumour area, whereas the remaining region T0 represents the normal area, as noted by an expert. An automatic technique predicts that the region P1 indicated in red is a tumour region, whereas the remaining region P0 is a normal region.

9. Limitations and challenges:

Although Swinunet is favoured due to its capacity to forecast out-of-training data, it still has significant limits and issues. It needs more processing resources because of the input pictures. More hyperparameters, as well as fine-tuning, are required. Training time is longer than for the Unext model. Swin-unet Networks are slower than other networks because they learn from all available information and are significantly more sophisticated. To address the processing power issue, we must use resources such as GPU and TPU, and because the merged model is a hybrid of the two models, it faces from the same issues.

10. Experimental setup:

The main aim is to conduct tumour segmentation from MRI scan pictures. First, we load all of the essential libraries. Then we'll extract and import our dataset. Because we couldn't train, we transformed all of the.mat files to numpy arrays and separated them into photos and tumour mass images. We made changes to the dataset to improve accuracy. We then finished preparing our dataset and moved on to data visualisation. We then began working on our models by importing various functions and dependencies from Tensorflow, Keras, and other libraries. However, the model requires a lot of RAM and a strong GPU machine during the training stage, therefore we trained our model using Google collab. However, during the testing stage, the system consumed more RAM and broke frequently due to inability to manage the dataset, which was largely images. Then we purchased Colab Pro in order to gain HIGH-RAM.

11. Discussion:

We are comparing our models because we are utilising two separate models and combining them to create a merged model. As a result, we wish to demonstrate that merged and ensemble models use less time and computing resources than a completely new architecture or model. The accuracy and graphs that we obtained are provided below.

S.NO	MODEL	Training Loss	Validation Loss	Training Dice Score	Validation Dice Score
1	UNEXT	0.0109	0.0429	0.7883	0.5184
2	SWIN-UNET	0.0085	0.0490	0.8323	0.5324
3	MERGED MODEL	0.0069	0.0313	0.8859	0.6799

Figure 15. Comparitive table of models

All three of our models are overfitting and unable to generalise to fresh data. The model performs very well on training data but poorly on fresh data in the validation set. The validity loss eventually diminishes but then begins to grow again. One significant cause of this occurrence is because

the model is either too sophisticated for the data or has been trained for an extended length of time. When the loss is low and stable, training can be stopped early, which is commonly referred to as early stopping. One of the various ways used to prevent overfitting is early stopping.

12. Conclusion:

In this project, we utilised two separate models from two different research papers and implemented them in our own method. We then attempted to develop a merged model utilising the models that were trained on the same dataset to determine which performed better in terms of accuracy. After applying our recommended models, we were able to successfully train and present our segmentation. We were able to acquire greater accuracies for the combined model as compared to the Swin-unet and Unext models, as predicted. In the future, we will aim to optimise and train the model using other datasets to see how it performs, as well as test new similar strategies and create models in such a manner that they provide greater accuracy with less processing resources.

References

[1] Cao H, Wang Y, Chen J, Jiang D, Zhang X, Tian Q, Wang M. Swin-unet: Unet-like pure transformer for medical image segmentation. arXiv preprint arXiv:2105.05537. 2021 May 12.

[2] Valanarasu, Jeya Maria Jose, and Vishal M. Patel. "UNeXt: MLP-based Rapid Medical Image Segmentation Network." arXiv preprint arXiv:2203.04967 (2022).

[3] Balwant MK. A Review on Convolutional Neural Networks for Brain Tumor Segmentation: Methods, Datasets, Libraries, and Future Directions. IRBM. 2022 May 13

[4] Long J, Shelhamer E, Darrell T. Fully convolutional networks for semantic segmentation. InProceedings of the IEEE conference on computer vision and pattern recognition 2015 (pp. 3431-3440).

[5] Ronneberger O, Fischer P, Brox T. U-net: Convolutional networks for biomedical image segmentation. InInternational Conference on Medical image computing and computer-assisted intervention 2015 Oct 5 (pp. 234-241). Springer, Cham.

[6] Chen J, Lu Y, Yu Q, Luo X, Adeli E, Wang Y, Lu L, Yuille AL, Zhou Y. Transunet: Transformers make strong encoders for medical image segmentation. arXiv preprint arXiv:2102.04306. 2021 Feb 8.

[7] Z. Zhou, M. Rahman Siddiquee, N. Tajbakhsh, and J. Liang, "Unet++: A nested u-net architecture for medical image segmentation." Springer Verlag, 2018, pp. 3–11.

[8] Ye L, Rochan M, Liu Z, Wang Y. Cross-modal self-attention network for referring image segmentation. InProceedings of the IEEE/CVF conference on computer vision and pattern recognition 2019 (pp. 10502-10511).https://slate.com/technology/2016/06/the-peculiar-endurance-of-the-physical-signature.html.

[9] Lian, D., Yu, Z., Sun, X., Gao, S.: As-mlp: An axial shifted mlp architecture for vision. arXiv preprint arXiv:2107.08391 (2021)

[10] Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. IEEE Transactions on Pattern Anal- ysis and Machine Intelligence, 2017.

[11] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. In International Conference on Learning Repre- sentations, 2015.

[12] A. Hatamizadeh, D. Yang, H. Roth, and D. Xu, "Unetr: Transformers for 3d med- ical image segmentation," 2021.

[13] K. S. P. J. M.-H. K. Isensee F, Jaeger PF, "nnu-net: a self-configuring method for deep learning-based biomedical image segmentation," Nat Methods, vol. 18(2):203- 211, 2021

[14] Çiçek Ö, Abdulkadir A, Lienkamp SS, Brox T, Ronneberger O. 3D U-Net: learning dense volumetric segmentation from sparse annotation. InInternational conference on medical image computing and computer-assisted intervention 2016 Oct 17 (pp. 424-432). Springer, Cham.

[15] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)