Multi-modal MRI-based Brain Tumor Segmentation

Majid Ahmed
Student, American University of Sharjah Sharjah, UAE
b00077868@aus.edu

Abstract—**The recent advances in parallelized computational power and machine learning gave rise to many machine learning applications in the medical field. One such application is the use of deep neural networks for segmentation. In this work, various machine-learning models are used to perform brain tumor segmentation using the BRATS2021 dataset. The dataset used is composed of 1251 multi-modal brain MRI scans with their corresponding masks that divide the MRI scan volume into 4 different segments. State-of-the-art models based on transformers and convolutional neural networks are trained and validated on 2D slices extracted from the dataset, and the results indicate that the SWIN-UNET transformer-based model performs the best using 2D axial slices and that the base UNET model performs comparably. Furthermore, the results suggest that utilizing different models trained on different anatomical 2D slices to create an ensembled segmentation model can increase the robustness of segmentation.**

Index Terms—**Brain tumor, MRI, image segmentation, axial, coronal, sagittal, UNET, TRANS-UNET, ATTEN-UNET, SWIN-UNET, PSPNET, LinkNet, FPN, BRATS2021, 3D segmentation**

I. Introduction

Magnetic Resonance Imaging (MRI) is a non-ionizing biomedical imaging modality used for numerous clinical applications. One such application is brain imaging to check for lesions or tumors. MRI scans are created by applying a strong static magnetic field, that can go up to 3 Tesla, to a patient such that most of the Protons of the body part being imaged are aligned in the same direction as the magnetic field. Further, a secondary time-varying Radio Frequency (RF) pulse is then applied periodically with a controlled frequency to change the protons' alignments temporarily to become transverse to the strong magnetic field. The RF pulse has a repetition time (TR) between consecutive pulses, and an echo time (TE) between each pulse to sample the signals generated by the protons as they are aligned back to the strong magnetic field. The time it takes for the protons to return to their initial alignment after the RF pulse is turned off is known as the T1-relaxation time. In addition, auxiliary gradient coils are used to create a spatially varying magnetic field to enable 3D brain imaging. Lastly, image reconstruction algorithms are then used to reconstruct a 3D image made up of voxels from the sampled signals. Measuring the T1-relaxation times results in t1-weighted MRI scans where tissues with short T1 relaxation times appear bright, and tissues with long T1 relaxation times appear dark. Furthermore, a contrast agent that contains gadolinium can be administered to the patient to shorten the T1 relaxation time and increase the contrast of lesions such that a T1-CE weighted scan is created.

On the other hand, the transverse relaxation time is known as the T2-relaxation time. A similar procedure with longer TR and TE times helps in generating t2-weighted MRI scans. The tissues with long T2 relaxation times appear bright in the scan, while tissues with short T2 relaxation times appear dark. T2-weighted scans require longer TR and TE relative to t1-weighted scans. Furthermore, fluid-attenuated inversion recovery (FLAIR) is a novel MRI modality in which the signals from free water protons are suppressed by applying two RF pulses [1]. An inversion Pulse flips all tissue protons by 180 degrees to cancel out signals from free water protons. An additional RF pulse that is known as the excitation pulse is applied to rotate the protons by 90 degrees to orient them transversely relative to the static magnetic field. The time between the inversion and excitation pulses is known as the inversion time, and it is optimized to be roughly equal to the T1 relaxation time of tissue-bound protons. Thus, in a flair-weighted MRI scan, tissues with high water content appear dark, while tissues with low water content appear bright. Hence, each MRI modality extracts varying information about the brain according to how the scan is acquired. The availability of such multi-modal Brain MRI scans plays a crucial role in training machine learning models capable of diagnosing patients and assessing the severity of cancerous growth inside the brain.

II. Literature Review

A. *Vox2Vox*

In [2], Generative Adversarial Networks (GANs) were utilized to perform semantic segmentation on brain tumors. The work was inspired by the PIX2PIX model, a famous implementation of image-to-image translation using GANs. The main advantage of using GANS is that they add another layer of punishment to unrealistic-looking segmentation masks. The dataset utilized, the BraTs2020 contains 3D MRI brain scans from 369 patients. For each patient, four types of MRI scans were collected. The MRI scan volume for each modality was fixed at 240 × 240 × 155. The approach proposed was to treat the problem as a 3D multi-channel segmentation problem. Each channel corresponds to one of the 4 MRI modalities used. A smaller sub-volume of 128 × 128 × 128 × 4 was extracted from each MRI scan to reduce the memory demand, and various 3D augmentations were used to avoid over-fitting. The proposed VOX2VOX model takes in this sub-volume and performs 3D convolution for feature extraction followed by the encoding layer. After the encoding layer, a transpose 3D convolution operation is performed to reverse the feature extraction process and reset the volume size to the original one. Lastly, a discriminator is added to assess the quality of the segmentation mask generated by the generator. The final prediction model was created by forming an ensemble of multiple Vox2Vox models, and it was reported to have a dice score of 93.40%.

B. UNET

UNET is a convolutional neural network designed for biomedical imaging segmentation tasks [3]. The network architecture includes a contracting path to capture context and a symmetric expanding path for precise localization. To achieve precise localization, high-resolution features from the contracting path are combined with the upsampled output. A successive convolution layer can then learn to assemble a precise segmentation mask based on this information. One important modification in the UNET architecture is that the upsampling part has a high number of feature channels to allow the network to propagate context information to higher-resolution layers. As a result, the expansive path is more or less symmetric to the contracting path and results in a U-shaped architecture. The network does not utilize fully connected layers, and the segmentation map only contains the pixels for which the full context is available in the input image.

C. Atten-UNET

Islam et al. [4] proposed a novel 3D UNET with the addition of an attention mechanism for MRI brain tumor segmentation. Attention is a transformer used to improve the performance of models designed to work with sequential or spatial data. This enables machine learning models to selectively focus on some relevant parts of the input data sequence whilst ignoring the rest. It is reported that the model achieved promising segmentation results with the 3D attention UNET architecture outperforming the original UNET model. The quantitative results for the BraTS 2019 testing set showed a mean Dice score of 77.8%, 86,89%, and 77.71% for the Enhancing Tumor (ET), Whole Tumor (WT), and Tumor Core (TC) sub-regions of the tumor, respectively.

D. Trans-UNET

Chen et al. [5] proposed a UNET-like neural network with the addition of transformers. The U-Net architecture has been widely used for medical image segmentation, but its convolution-based operations have limitations in explicitly modeling long-range dependencies. On the other hand, Transformers have innate global self-attention mechanisms but lack the low-level details necessary for medical imaging. Hence, TransUNet combines both Transformers and U-Net to provide accurate medical image segmentation. TransUNet establishes self-attention mechanisms from the perspective of sequence-to-sequence prediction, leveraging the Transformer's ability to encode tokenized image patches from a CNN feature map as input sequences for extracting global contexts. The encoder encodes the image representation through multi-head self-attention, followed by a fully connected feed-forward network. Layer normalization is applied before each sub-layer. The decoder then upsamples the encoded features, combining them with high-resolution CNN feature maps to enable precise localization similar to UNET. To evaluate the performance of TransUNet, various medical image segmentation tasks were tested, including multi-organ segmentation and cardiac segmentation. The results show that TransUNet outperforms other CNN-based self-attention methods, U-Net, and other CNN-based methods.

E. Swin-Unet

In [6], a pure transformer-based neural network is proposed for medical image segmentation. The transformer architecture is used to attend to different parts of an input sequence parallelly in contrast to recurrent neural networks in which an input sequence is processed sequentially. Further, this parallelization enables the model to learn dependencies between different areas in the MRI scans more effectively. The proposed Swin-UNET contains an encoder, bottleneck, decoder, and skip connections. The encoder is used to transform the input images into sequence embeddings by splitting up the images into non-overlapping 4x4x3 patches. The patches then undergo a linear embedding process to prepare them to be passed on to the Swin transformer blocks. A symmetrical decoding block is used to reverse the process in addition to the skip-ahead connections such that the output segmentation mask is up-sampled to the same size as the input image. Skip connections are used to utilize the shallow features extracted at the early stages of the neural network along with the deeper features extracted at the later stages of the model to minimize the spatial information loss that results from the down-sampling and up-sampling processes. The model was trained and validated using the Synapse multi-organ segmentation dataset and it outperformed other models such as UNET, Atten-UNET, and Trans-UNET.

III. Methodology

Various machine learning models will be tested using the BRATS2021 dataset [7]. The dataset contains 1251 multi-modal scans with a 240x240x155x4 multi-channel volume where each channel corresponds to one of the four MRI modalities used. However, the t1-weighted scans were ignored such that the new input shape is reduced to 240x240x155x3. The dataset was then divided randomly to have a training subset of 1000 scans and a validation subset of 251 scans. Due to the limited available computational power, the 3D segmentation task will be addressed as a 2D segmentation task. 100 slices out of the 155 slices were taken as most of the ignored slices had minimal information. The slices were taken from the axial plane. In addition, each image was further cropped to have a shape of 128x128x3 to decrease the computational requirements. Each mentioned model was trained for 20 epochs using the categorical cross-entropy as a loss function. The metric used to quantify the performance of a machine learning model doing a semantic segmentation task is the dice loss. The dice loss quantifies the percentage overlap between the true and predicted masks such that it has a value between 0 and 1.

The preprocessed image dataset was utilized in 3 different comparative studies using different machine-learning models. In the first study, four segmentation models were tested on the BRATS2021 dataset. The first model used is UNET as discussed above. The second model trained is the Pyramid Scene Parsing Network (PSPNET) proposed in [8].PSPNET uses a pyramid pooling module that captures contextual information at multiple scales by dividing the input feature maps into fixed-size sub-regions and applying max-pooling operations to each sub-region. This allows the network to capture information at different levels of abstraction, from fine-grained details to global context. Furthermore, the third model is the Feature Pyramid Network (FPN) proposed in [9]. FPN improves on PSPNet by adding connections between different levels of the feature pyramid, allowing information to flow up and down the pyramid. This enhances the network's capability of utilizing multi-scale contextual information to achieve more accurate and robust object detection and segmentation. Furthermore, the fourth model used is LinkNET [10] which has a somewhat similar structure to UNET. The model uses a set of encoding and decoding blocks in addition to the use of skip connections.

The second study was performed to compare the performance of different UNET-based transformer models [11]. UNET, Trans-UNET, Atten-UNET, and SWIN-UNET were all trained similarly on the 2D axial MRI slices to perform the segmentation task. Lastly, the performance of UNET and SWIN-UNET using 2D slices from different anatomical planes, i.e. axial, coronal, and sagittal planes, was studied to quantify the dependency of the 2D segmentation performance on the anatomical plane from which the slices are extracted.

IV. Results

A. Basic segmentation models

PSPNET, FPN, Link-NET, and UNET are all high-level neural network architectures used for semantic segmentation. Hence, they can be built using well-established CNN architectures as backbones for feature extraction. The backbone architecture used to create each segmentation model was varied using a smaller subset of the dataset to maximize each model's performance. The initial results indicated that utilizing a SERESNEXT101 backbone produced the best performance for both UNET and PSPNET. On the other hand, MobileNETV2 performed the best for the FPN model, and EfficientnetB4 was used as a backbone for Link-NET.

The training loss for each model is summarized in Fig.1. The training results indicate that the FPN model has the best training performance. However, the validation data suggest that this apparent performance advantage is only due to the model over-fitting to the training dataset to some extent as suggested by the validation results in Fig.2. In addition, the validation dice coefficient corresponding to the enhancing, edema, and necrotic tumor regions is reported separately since the overall dice score does not provide enough information about the models' performance. The results indicate that the UNET model performs the best by a significant margin in correctly segmenting both necrotic and enhancing tumor regions while maintaining a small lead in segmenting edema regions correctly.

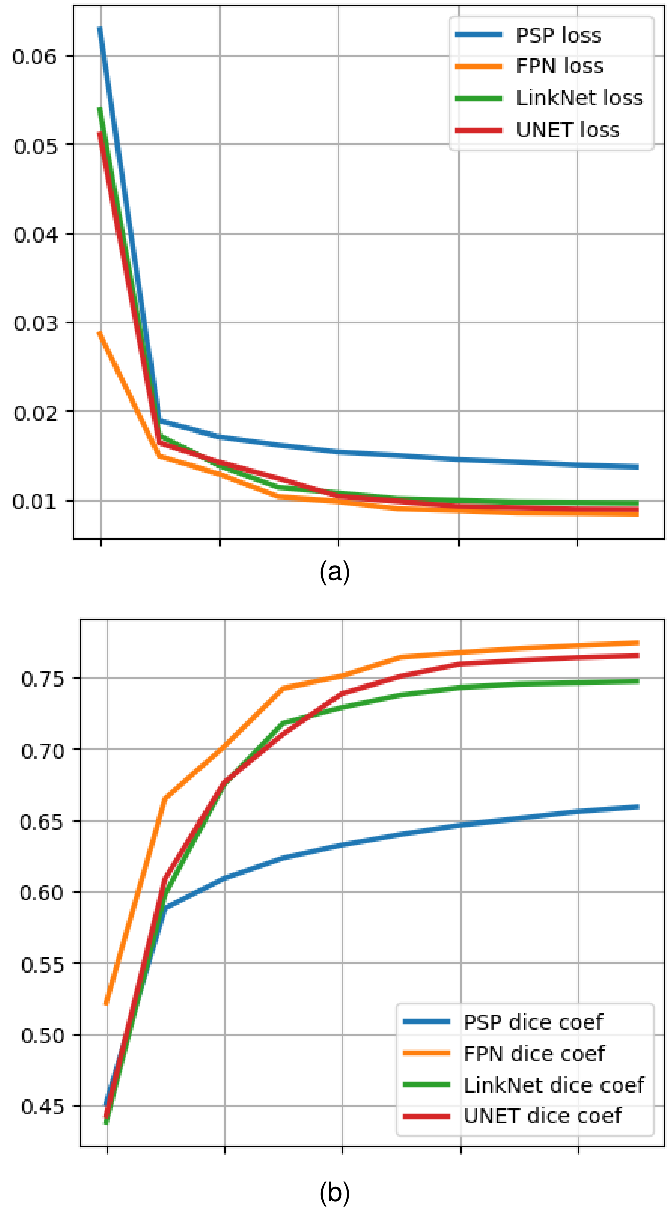


Fig. 1. (a) base models training loss, and (b)overall dice score vs epochs no..

B. UNET based architectures

For the second study, different UNET-based models were trained for 20 epochs each. The training results summarized in Fig.3 indicate that the base UNET model has the best training performance as it has the lowest loss and highest overall dice coefficient. However, the validation results plotted in Fig.4 indicate that the SWIN-UNET model performs significantly better in segmenting the edema region of the brain tumors. Further, the Swin-UNET model matches the performance of the base UNET model in segmenting the enhancing tumor

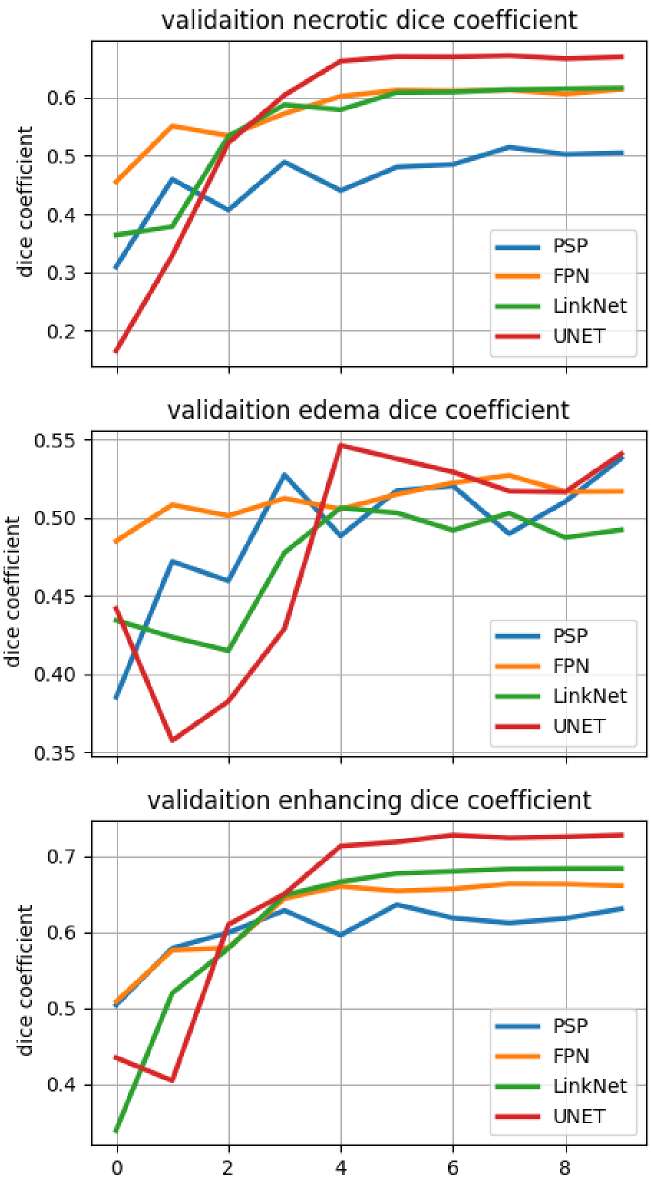
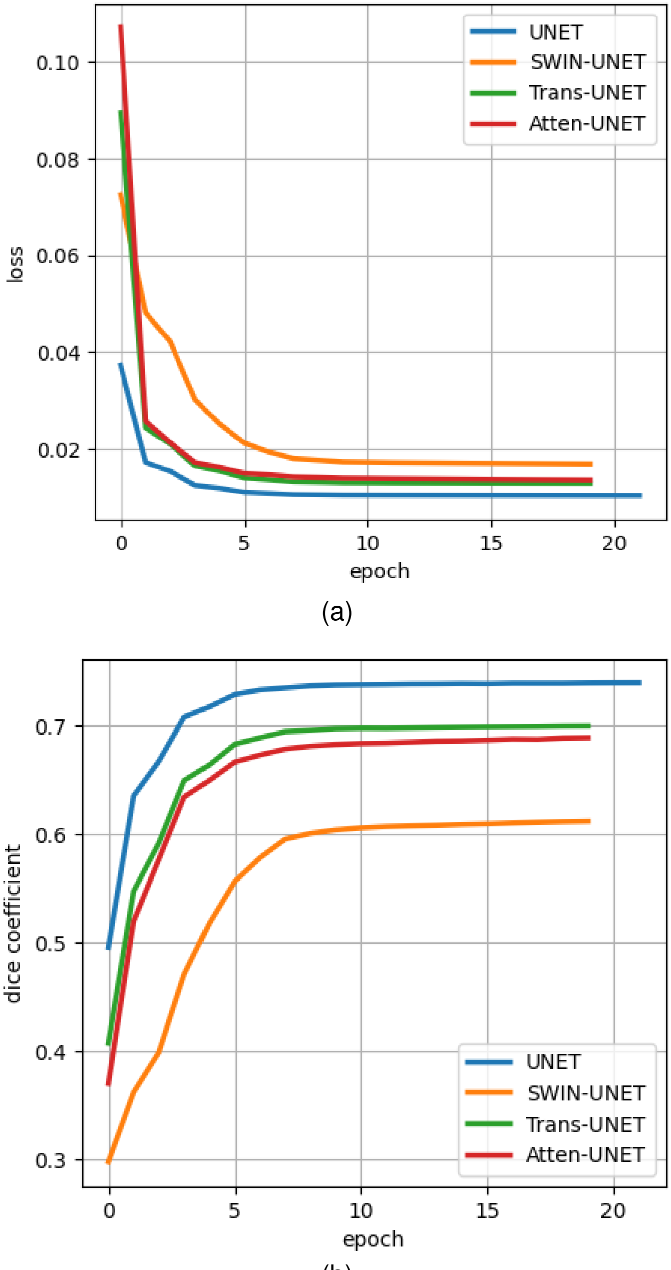


Fig. 2. base models validation dice coefficients.

region, but the model struggles to identify the necrotic tumor segment as it has poorer performance when compared to the UNET model. A potential explanation is that the transformer-based models struggle to identify the necrotic tumor region since it usually has the smallest volume in the MRI scan. Thus, utilizing a global attention approach decreases the performance in detecting the small necrotic tumor region as extracting any relevant features to it becomes significantly harder. This can be solved by utilizing a more localized transformer approach or introducing class weighting to the loss function. Nevertheless, the Swin-UNET model maintains an almost similar performance on both the validation and training sets whilst having a huge improvement in segmenting edema tumor regions.



(a)

(b)

Fig. 3. (a) UNET-based models training loss, and (b)overall dice score vs epochs no..

C. *Revised Performance with different anatomical slices*

For the final study, the plane from which the 2D slices are taken was varied to quantify how the segmentation models' performance depends on it. In the previous sections, all models were trained using 2D slices extracted from the axial plane of the MRI scans. The same UNET and Swin-UNET models were retrained using 2D slices from the sagittal and coronal planes. The validation results summarized in Fig.**??** indicate that training the base UNET model on coronal 2D slices significantly enhanced its capability of segmenting the edema tumor region. However, this increase is paired with some per-

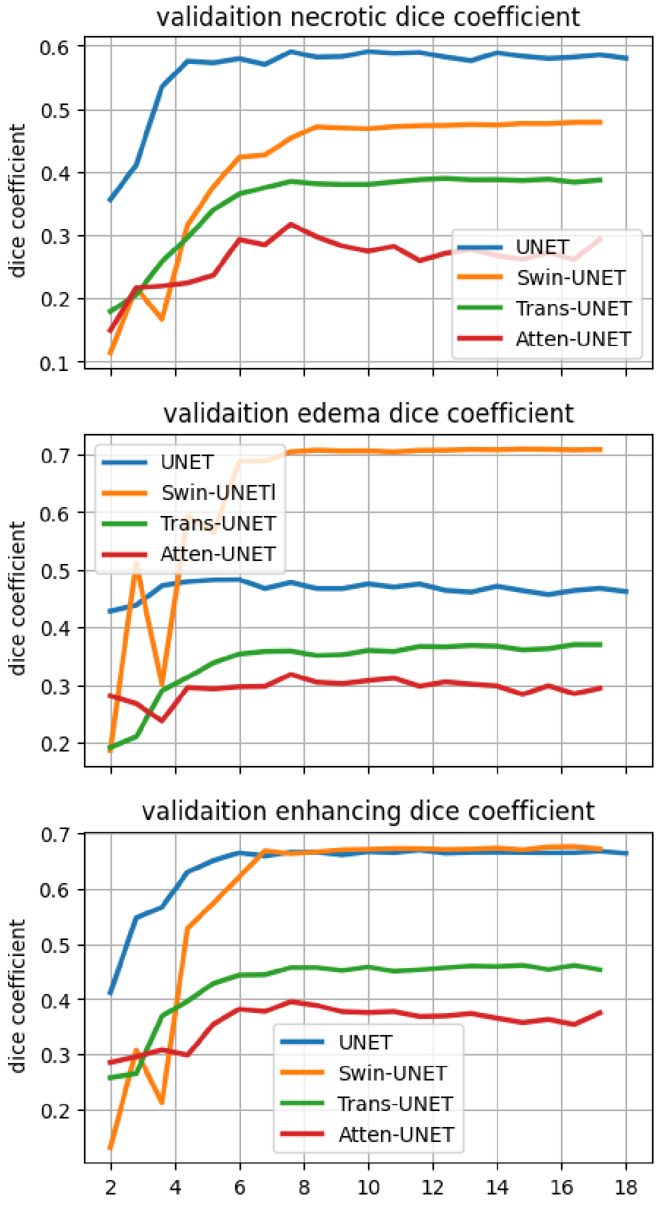


Fig. 4. base models validation dice coefficients.

formance degradation in segmenting the other tumor regions. On the other hand, using coronal slices negatively affects the performance of the Swin-UNET model in segmenting all three tumor regions significantly. Consequently, the anatomical plane from which the 2D slices are taken plays a major role in dictating the performance of the segmentation model.

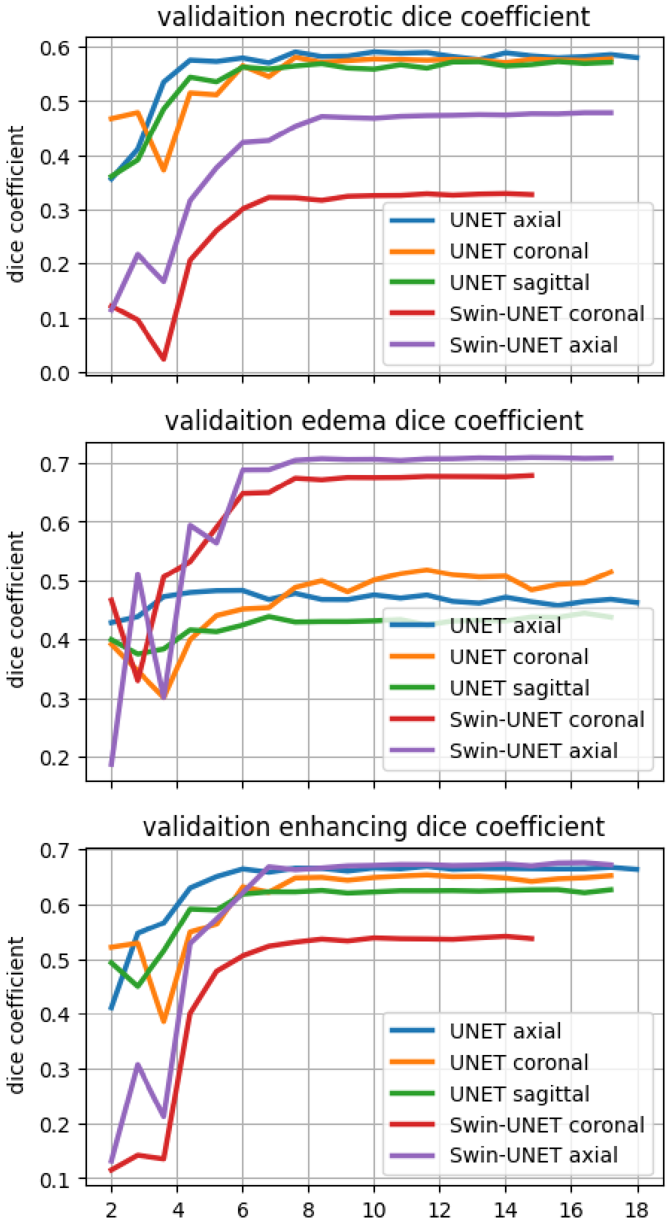


Fig. 5. validation dice coefficients of UNET models trained on slices from different 2D anatomical planes.

V. Conclusion

Different segmentation models were trained on 2D slices obtained from the BRATS2021 dataset, and their performance was compared using the dice coefficient metric. The experimental results suggest that the best-performing models were Swin-UNET and UNET when trained on 2D slices obtained from the axial plane. In addition, the results indicate that examining the performance of the other models with slices from different anatomical planes can potentially produce a partial enhancement in some of the models' performance. An alternative approach that could have been studied was the use of GANs which could have enabled enhancing the resolution of the images in addition to creating the segmentation masks. In addition, most of the utilized transformer-based models were used in their simplest forms with a small number of convolutional filters and transformer layers to enable training using the limited computational power. Hence, the Atten-UNET and Trans-UNET models had the poorest performance out of all the utilized models.

Although the experimental results suggest that the best-performing models were the UNET and Swin-UNET segmentation models, it must be noted that the approaches studied are of only single models. A more realistic approach is to utilize ensembles of different models trained for varying numbers of epochs and with different hyper-parameters to increase the robustness of the segmentation process. Furthermore, all of the trained models for this project have much lower performance compared to what is reported in the literature. This can be attributed to the use of 2D slices instead of 3D volumes since it reduces the amount of spatial information available to extract features from. In addition, the performance of the trained models could have been increased by utilizing additional augmentation techniques such as changing the brightness of the scans, plastic deformation, shearing, or others to generate new input data from the existing dataset. However, this was not pursued due to the limited computational power available. Nevertheless, a more appropriate approach would have been to create an ensemble of segmentation models trained on the 3D MRI scans such that the full 3D context of the scan is maintained. The tumor segmentation could then be done using a majority vote technique, or by assigning each model in the ensemble a weight according to its respective performance in segmenting the three tumor regions. In the context of 2D brain tumor segmentation, the segmentation models can be trained on different anatomical planes, and the final 3D tumor mask can be reconstructed from the ensemble's 2D output slices.

Acknowledgment

N/A

References

[1] A. Murphy and F. Gaillard, "MRI sequences (overview)," jun 2015.

[2] M. D. Cirillo, D. Abramian, and A. Eklund, "Vox2vox: 3d-gan for brain tumour segmentation," Mar. 2020.

[3] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," May 2015.

[4] M. Islam, V. VS, V. J. M. Jose, N. Wijethilake, U. Utkarsh, and H. Ren, "Brain tumor segmentation and survival prediction using 3d attention unet," Apr. 2021.

[5] J. Chen, Y. Lu, Q. Yu, X. Luo, E. Adeli, Y. Wang, L. Lu, A. L. Yuille, and Y. Zhou, "Transunet: Transformers make strong encoders for medical image segmentation," Feb. 2021.

[6] H. Cao, Y. Wang, J. Chen, D. Jiang, X. Zhang, Q. Tian, and M. Wang, "Swin-unet: Unet-like pure transformer for medical image segmentation," May 2021.

[7] U. Baid, S. Ghodasara, S. Mohan, M. Bilello, E. Calabrese, E. Colak, K. Farahani, J. Kalpathy-Cramer, F. C. Kitamura, S. Pati, L. M. Prevedello, J. D. Rudie, C. Sako, R. T. Shinohara, T. Bergquist, R. Chai, J. Eddy, J. Elliott, W. Reade, T. Schaffter, T. Yu, J. Zheng, A. W. Moawad, L. O. Coelho, O. McDonnell, E. Miller, F. E. Moron, M. C. Oswood, R. Y. Shih, L. Siakallis, Y. Bronstein, J. R. Mason, A. F. Miller, G. Choudhary, A. Agarwal, C. H. Besada, J. J. Derakhshan, M. C. Diogo, D. D. Do-Dai, L. Farage, J. L. Go, M. Hadi, V. B. Hill, M. Iv, D. Joyner, C. Lincoln, E. Lotan, A. Miyakoshi, M. Sanchez-Montano, J. Nath, X. V. Nguyen, M. Nicolas-Jilwan, J. O. Jimenez, K. Ozturk, B. D. Petrovic, C. Shah, L. M. Shah, M. Sharma, O. Simsek, A. K. Singh, S. Soman, V. Statsevych, B. D. Weinberg, R. J. Young, I. Ikuta, A. K. Agarwal, S. C. Cambron, R. Silbergleit, A. Dusoi, A. A. Postma, L. Letourneau-Guillon, G. J. G. Perez-Carrillo, A. Saha, N. Soni, G. Zaharchuk, V. M. Zohrabian, Y. Chen, M. M. Cekic, A. Rahman, J. E. Small, V. Sethi, C. Davatzikos, J. Mongan, C. Hess, S. Cha, J. Villanueva-Meyer, J. B. Freymann, J. S. Kirby, B. Wiestler, P. Crivellaro, R. R. Colen, A. Kotrotsou, D. Marcus, M. Milchenko, A. Nazeri, H. Fathallah-Shaykh, R. Wiest, A. Jakab, M.-A. Weber, A. Mahajan, B. Menze, A. E. Flanders, and S. Bakas, "The rsna-asnr-miccai brats 2021 benchmark on brain tumor segmentation and radiogenomic classification," Jul. 2021.

[8] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," Dec. 2016.

[9] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," Dec. 2016.

[10] A. Chaurasia and E. Culurciello, "Linknet: Exploiting encoder representations for efficient semantic segmentation," dec 2017.

[11] Y. Sha, "Keras-unet-collection," https://github.com/yingkaisha/keras-unet-collection, 2021.

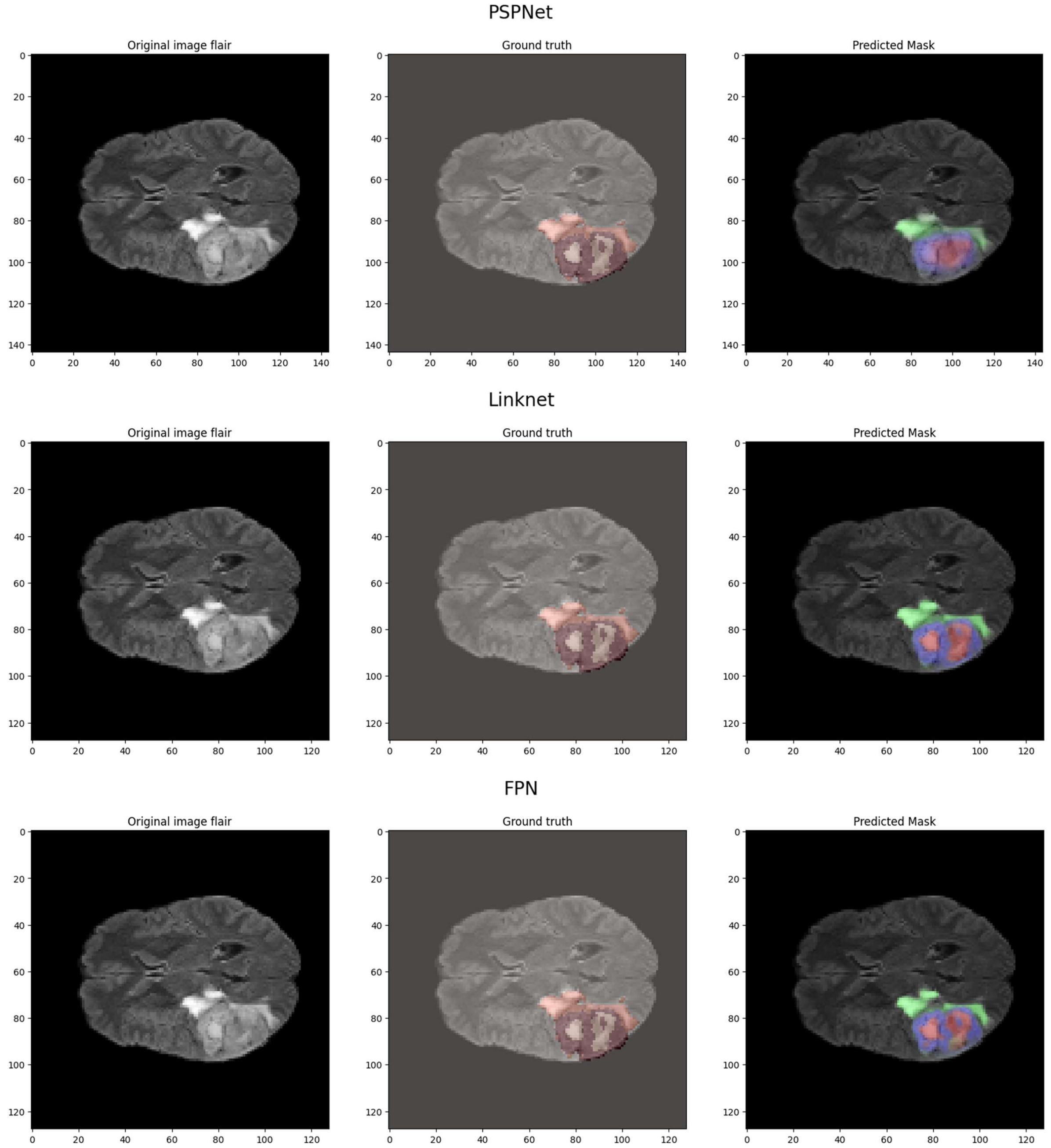


Fig. 6. different segmentation models comparison.

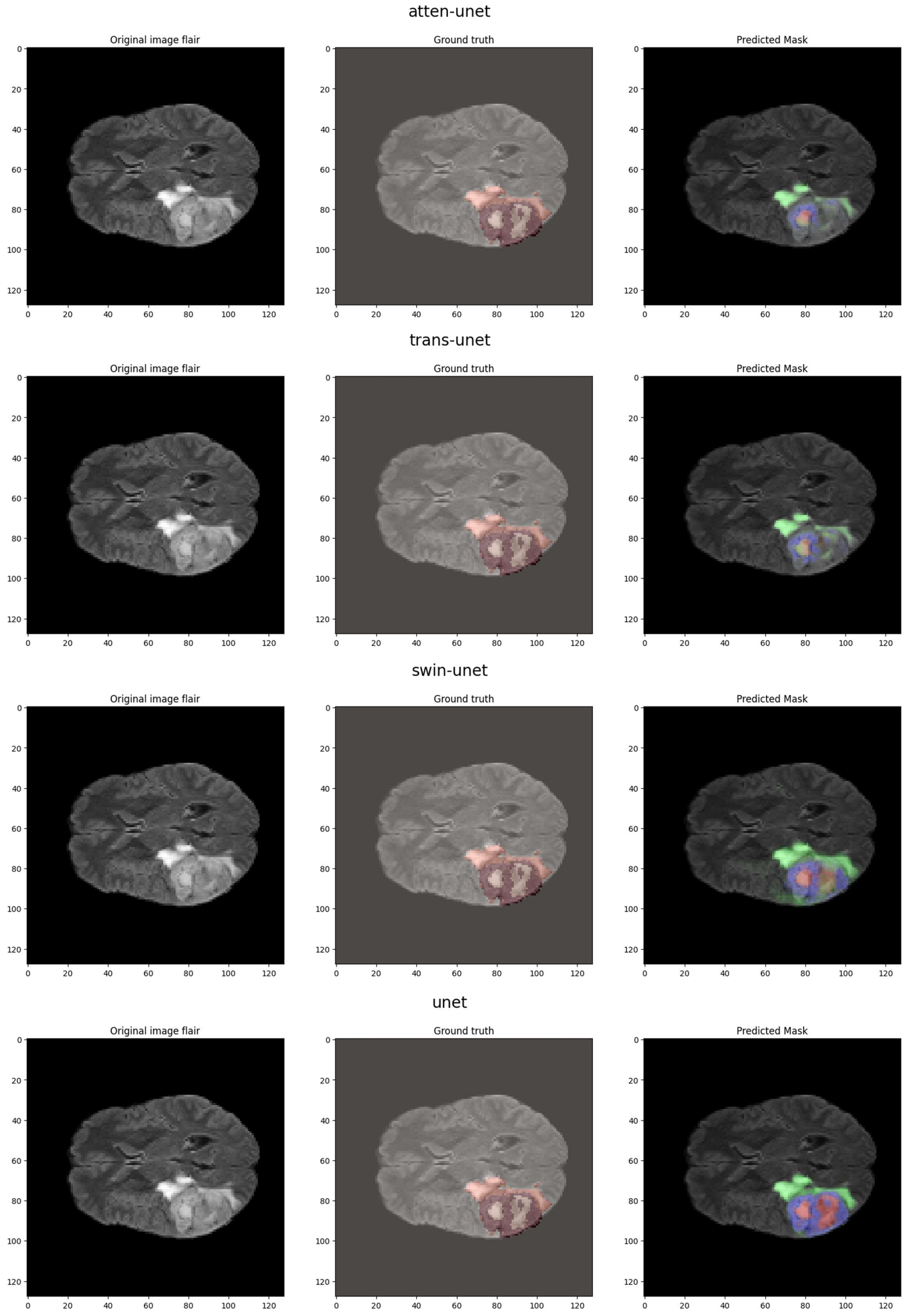


Fig. 7. different UNET-based models segmentation comparison.

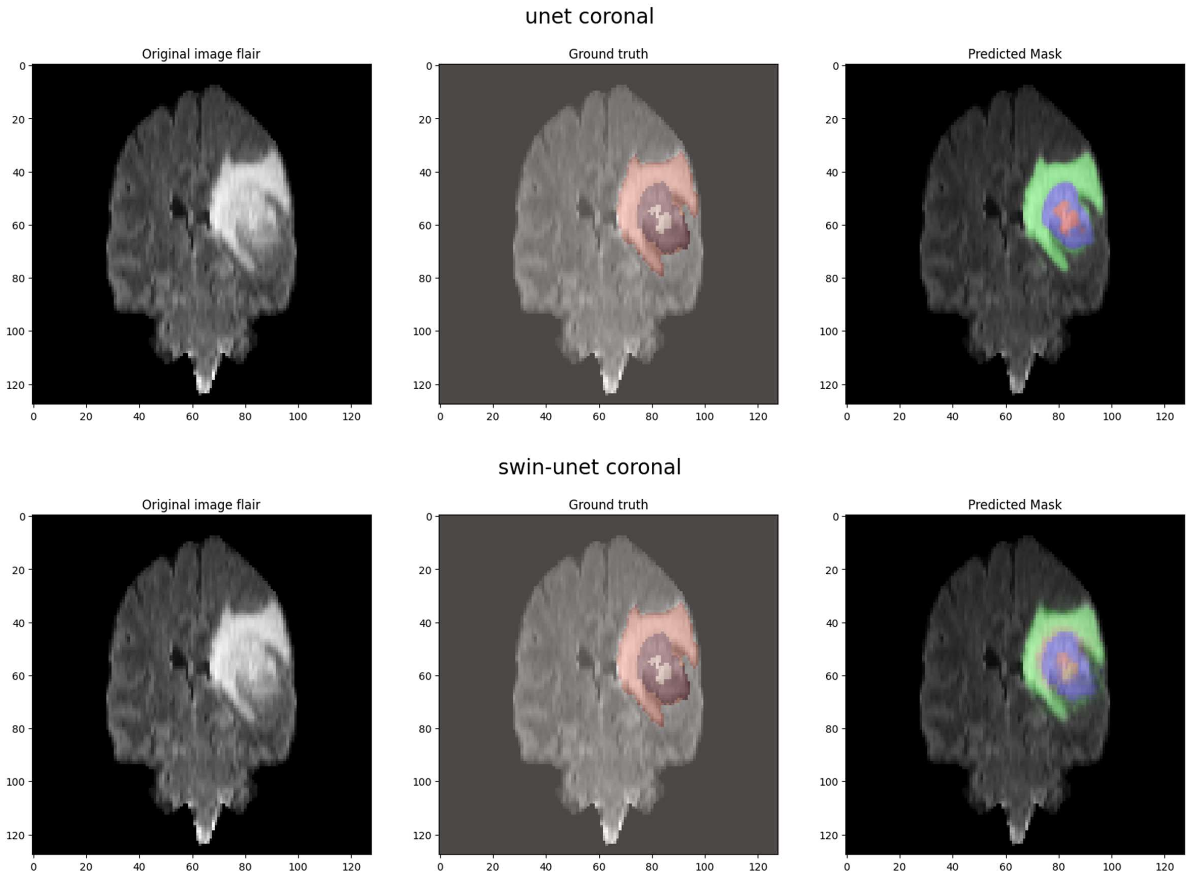


Fig. 8. UNET model segmentation comparison with coronal slices.