A Multi-scale Fusion Deep Learning Approach on Brain Tumor Segmentation

Yuke Zhang, Xinyi Hu, Seung Hee Lee

January 26, 2024

Abstract

The objective of the research is to propose a novel Multimodal and Multiscale model for the classification and segmentation of brain tumors on Magnetic Resonance Imaging (MRI) images. We carefully analyze the Brain Tumor Segmentation(BraTS) 2021 Dataset with the intention of leveraging these images for diagnostic purposes. Here is the link to our project repository: `https://github.com/cassielee04/smart_brains`. We found that the multi-modality model performed better than the single modality model for the who brain tumor segmentation task in BraTS.

Introduction

Detecting brain tumors in large medical images databases takes a lot of effort and time in manual clinical tasks. That's why Automated brain tumor segmentation and classification is important for medical diagnosis [1]. Considering a multiscale approach is effective extracting discriminant texture features of different kinds of tumors, we propose a deep convolutional neural networks with a multiscale feature extraction approach. Multiscale receptive fields on deep-feature maps is good at capturing the local and contextual object information[2]. The ultimate goal is to enhance the accuracy and effectiveness of MRI as a non-invasive, cost-effective tool for diagnosing, predicting growth, and treating brain tumors, underlining the critical role of automatic segmentation and classification in medical diagnostics. One way to enhance classification accuracy is by utilizing data from multiple modalities. We suggest that leveraging multi-modal imaging will improve tumor segmentation.

Related Work

In order to tackle the brain tumor classification task, both traditional machine learning techniques, such as Support Vector Machines (SVM) and Fisher kernels, and advanced deep

learning strategies using various architectures of Convolutional Neural Networks (CNN) are applied. CNN-based methods demonstrate considerable effectiveness, especially when enhanced with appropriate data augmentation techniques including geometric transformations, grayscale modifications, and other manipulations. In many Brain Tumor Segmentation application, segmentation are essential treated as classification by classifying pixels.

Díaz-Pernas et al. distinguish their approach by adeptly incorporating multiscale information, utilizing kernels of three predetermined sizes to extract and concatenate features concurrently. This method surpasses others by leveraging a nuanced understanding of how different scales of information contribute to more accurate tumor classification[1].

Another method is CNN with U-net by Mostafa et al. The U-Net architecture, renowned for its efficacy in image segmentation, forms the basis of the model when applying CNN with the U-Net sampling technique. This architecture is structured with convolutional and max-pooling layers for feature detection, upsampling layers to restore spatial resolution, and concatenation operations to merge features at different levels. Such a design equips the model to adeptly capture and interpret both the intricate, low-level details and the overarching, high-level characteristics of the input images, ensuring a thorough and nuanced understanding essential for accurate segmentation[3].

MRI scans provided hundreds of 2D images with high soft tissue contrast, and different MRI modalities highlight and contrast different medical information. T1 modality is mainly for healthy tissues. T2 images are proper to detect borders of edema regions. T1CE images highlight tumor borders and FLAIR images favor the detection of edema regions in Cerebrospinal Fluid. Consider the goal of detecting brain tumor, T1CE and FLAIR are favorable. Therefore, we will start with exploring T1CE and FLAIR first.

Proposed Work

Díaz-Pernas et al. suggested that using only the T1CE imaging modality is sufficient for tumor segmentation tasks. However, considering the unique characteristics of different MRI modalities and our project's goal to perform image segmentation on the BraTS dataset – which aims at a different segmentation task than Díaz-Pernas et al. – we plan to utilize both FLAIR and T1CE modalities. Our approach involves integrating multi-scale models (feature pyramids) for FLAIR and T1CE modalities and combining T1CE and FLAIR (T1CE-FLAIR) data before assessing the efficacy of using a single modality.

The study referenced [1] demonstrates a classification accuracy of 0.973 with the T1CE modality alone for differentiating tumor types. Nevertheless, the BraTS dataset challenges us to not only classify but also accurately identify edema—swollen areas caused by fluid accumulation. FLAIR images excel at revealing edema, implying their inclusion could

enhance our model's segmentation performance. By combining FLAIR with T1CE modalities, we expect not just better tumor identification accuracy but also superior segmentation of edema regions.

Inspired by the multi-scale feature extraction deep learning discussed in the referenced study on the U-Net architecture[4] (see Figure 1), we re-implement the same model architecture for the purpose of multi-scale feature extraction for each modality set independently on the brain image slice with id 78. Max-min normalization is applied on every slice, before pass into the U-Net. We also made the following modifications compared to the original U-Net. We incorporate the Batch Normalization method [5]. Given that the BraTS Dataset encompasses four distinct classification categories, there will be four channels corresponding to each classification class before the final layer. Subsequently, a softmax layer will be applied to these four channels to generate the final segmentation task. The number of input channels depends on the number of modalities defined for the model.

Loss function

Since Dice score is in general a better metric to measure picture similarity compared to accuracy, our model will use customized MultiClassDiceLoss instead of Cross-Entropy for loss function. MultiClassDiceLoss is defined as

$$\text{MultiClassDiceLoss}(P,T) = \left[\sum_{i=1}^{\text{n\_class}} 1 - \frac{2 \times |P_{i1} \cap T_{i1}|}{|P_{i1}| + |T_{i1}|}\right] / \text{n\_class},$$

where P represents the prediction, T represents the ground truth, and P_{i1} and T_{i1} represent the sets of voxels classified as the current category i in the prediction and ground truth. As explained in the Dataset section, our targets include four types of classification: Not tumor, Necrotic Core, Edema, and Enhancing. Therefore, our loss function is MultiClassDice over the four classes.

Considering the limited amount of data available from Kaggle, we also applied data augmentation methods. We experimented with three types of modality inputs—T1CE, FLAIR, and T1CE-FLAIR—both with and without data augmentation methods, resulting in a total of six models.

Data Augmentation

In previous literature, researchers have demonstrated that data augmentation can enhance model performance, mitigate overfitting, and act as a regularizer [6]. For our study, we implemented several augmentation techniques, including Gaussian noise addition with a small variance of 0.00001, horizontal and vertical flips, limited rotation of up to 10 degrees,

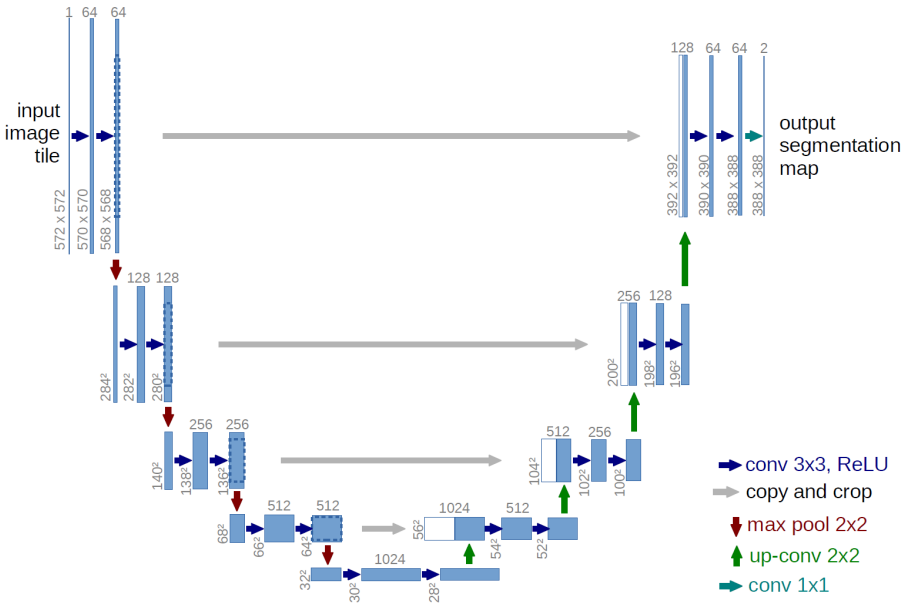


Figure 1: U-Net in [4]

small shifts without scaling, and small shearing within a range of -5 to +5 degrees along both x and y axes. The probability of applying each transformation is set to 0.5.

Datasets

We used the BraTS 2021 Dataset, part of the Brain Tumor Segmentation (BraTS) challenge series, is a comprehensive collection of multimodal magnetic resonance imaging (MRI) scans used for the segmentation of brain tumors, specifically gliomas. Gliomas are common types of primary brain tumors, which are categorized into low-grade gliomas (LGGs) and high-grade gliomas (HGGs), with the latter being more aggressive and having a poorer prognosis.We obtained data from the following Kaggle dataset: BRATS2021 Training and Validation.We have labeled image data for 340 subjects, each with 4 image modalities. We randomly selected 10 subjects as stored as unseen test data. We divided the rest of dataset into training and validation sets using an 80:20 split.

The ground truth segmentation image in the BraTS dataset segmented a brain scan into the following 4 categories: '0': Not tumor; '1': Necrotic/Core; '2': Edema; '3': Enhancing. The BraTS multimodal MRI scans dataset comprises a comprehensive collection of images available in NIfTI format (.nii.gz). It includes four types of MRI volumes: a) native T1-weighted (T1), b) post-contrast T1-weighted (T1CE), c) T2-weighted (T2), and d) T2 Fluid Attenuated Inversion Recovery (T2-FLAIR) volumes. These images were sourced from

multiple institutions using varied clinical protocols and scanners, enhancing the dataset's diversity and complexity.

The BraTS dataset serves as a benchmark for testing BT detection and segmentation algorithms within the research community [7]. It encompasses multiple MRI modalities such as FLAIR, T2-weighted, and T1-weighted with contrast, along with detailed annotations for various tumor regions, using binary masks to distinguish tumor tissue from the background.

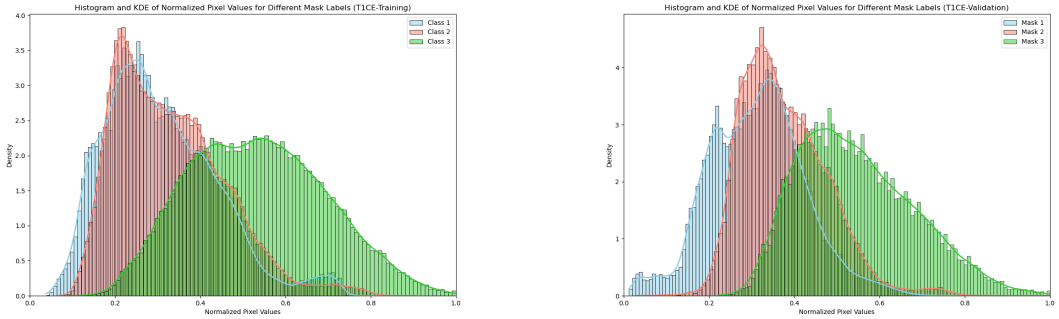


Figure 2: Histogram and KDE of Normalized Pixel Values for Different Mask Labels (T1CE-Training and T1CE-Validation)

Training and Validation Set Distribution T1CE

We plotted the distributions for the segmentation of the four categories mentioned above for each normalized brain image with slice ID 78. For every image, regardless of modality, we applied max-min normalization so that all the pixel value is between 0 and 1.

In the T1CE MRI scan datasets (see Figure 2), Class 1 of the training set shows a distribution skewed towards lower pixel values, denoting predominantly low-intensity areas, while Class 2 has a a similar but distinguishable distribution compared to Class 1. Class 3's distribution peaks at higher pixel values, typical of areas with higher intensity in the scans. The validation set echoes these characteristics, with Class 1 resembling the low-intensity skew of Class 1, Class 2 showing a slightly broader intensity range than its training counterpart, and Class 3 presenting a wider spread across mid to high intensities.

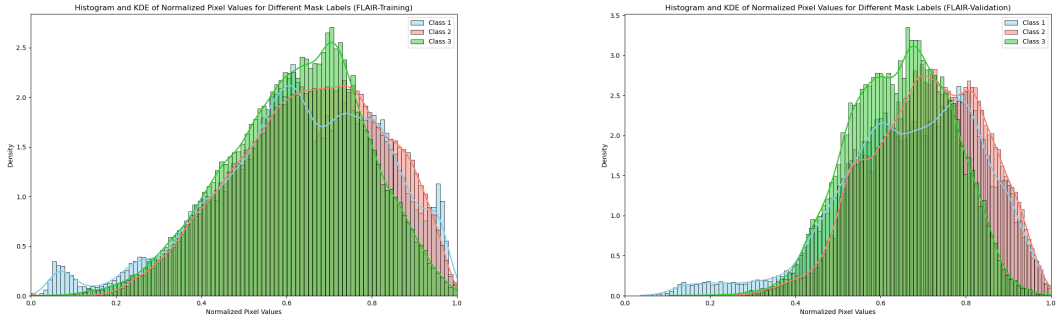


Figure 3: Histogram and KDE of Normalized Pixel Values for Different Mask Labels (FLAIR-Training and FLAIR-Validation)

Training and Validation Set Distribution FLAIR

For the FLAIR MRI scan datasets (see Figure 3), the training set distribution demonstrates that besides Class 1 has a small and left-skewed peak for lower pixel values, all three classes have a broad and right-skewed distribution, suggesting brighter regions within the scans. FLAIR distributions for three classes are more overlapped compared to T1CE. Similar pattern is observed in the validation set as well. The distributions of the three classes in FLAIR overlap more significantly as compared to those in T1CE. This overlap suggests that it might be more challenging for the some models to distinguish between the classes when analyzing FLAIR images. The advantage of using U-Net lies in its ability to leverage the spatial information contained within the image.

Evaluation

Segmentation is achieved by classifying each pixel in the MRI scan image into one of four categories:

1. '0': Not tumor;
2. '1': Necrotic Core;
3. '2': Edema;
4. '3': Enhancing.

Our model's performance will be evaluated using metrics such as accuracy, precision, sensitivity (true positive rate), specificity (true negative rate), F1 score, and Dice score [7],[3].

Precision, sensitivity, specificity, and F1

Precision, sensitivity, specificity, and F1 Score are defined as follows:

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}}$$

$$\text{Sensitivity} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}} \quad \text{and} \quad \text{Specificity} = \frac{\text{True Negatives}}{\text{True Negatives} + \text{False Positives}}$$

$$\text{F1 Score} = 2 \times \frac{\text{Precision} \times \text{Sensitivity}}{\text{Precision} + \text{Sensitivity}},$$

Dice Score

The Dice Similarity Index, a measure of similarity between two images, is defined as:

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

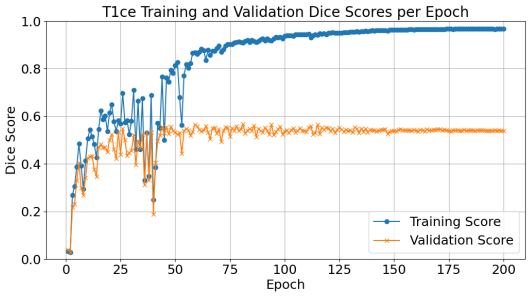
where P represents the prediction, T represents the ground truth, and P_1 and T_1 represent the sets of voxels classified as the current category in the prediction and ground truth, respectively.

Dice score is not only applied at the end for model evaluation. We also use the Dice score for model selection. We save the best model that predicts segmentation with the best average Dice score on classes '1', '2', and '3' on the validation dataset during the training process. The reason we do not include class '0' is that class '0' represents the background class, and the Dice score for class '0' is usually much higher than for other classes.

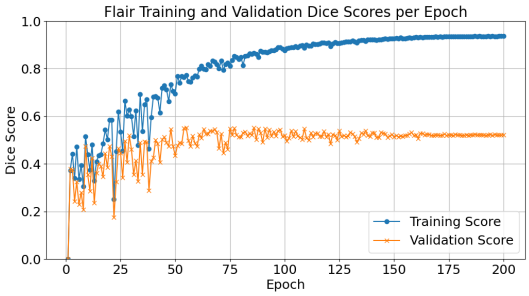
Result

In this section, we present a comparative analysis of model performance with and without the implementation of data augmentation techniques. We will demonstate impact of augmentation on the model's ability to accurately segment tumor sub-regions within MRI scans.

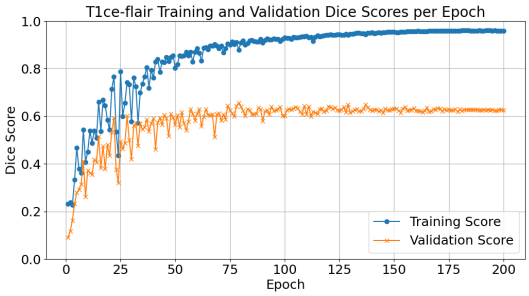
Without Data Augmentation



(a) T1CE



(b) FLAIR

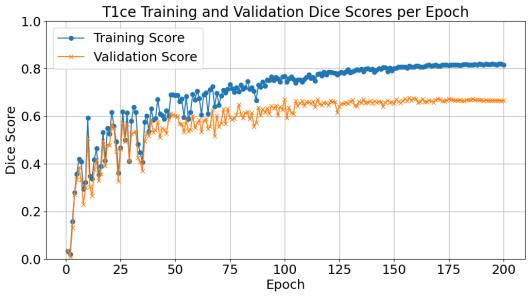


(c) T1CE-FLAIR

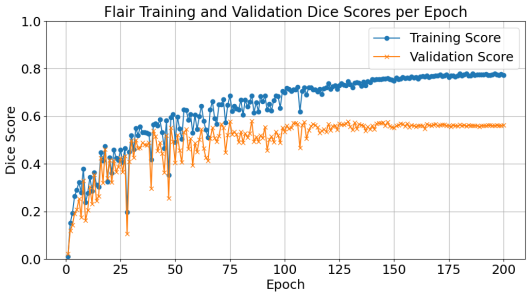
Figure 4: Model performance without data augmentation

The set of graphs in (see Figure 4) shows the training and validation Dice scores for models trained on FLAIR, T1CE, and combined T1CE-FLAIR MRI scans without the use of data augmentation. In the FLAIR and T1CE models, the training scores demonstrate a steady increase and eventually plateau. Conversely, the validation scores remain relatively constant throughout the epochs, suggesting a lack of generalization from the training data to the validation data. This performance plateau in validation Dice scores indicated an overfitting to the training data. The combined T1CE-FLAIR model also shows a similar pattern, with the training Dice score achieving high values while the validation score plateaus at a lower level, reinforcing the need for techniques like data augmentation to bridge the gap between training and validation performance and enhance the model's ability to generalize.

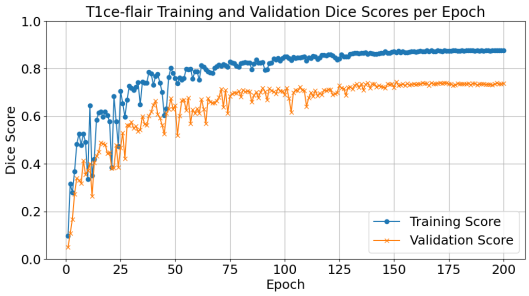
Data Augmentation



(a) T1CE



(b) FLAIR



(c) T1CE-FLAIR

Figure 5: Model performance with data augmentation

The set of graphs (see Figure 5) display the trajectory of Dice scores across training and validation phases for models trained on augmented FLAIR, T1CE, and T1CE-FLAIR MRI data. The training scores for all models increase quickly in the initial epochs before leveling off, signifying a point of diminishing returns in learning. The validation scores, while showing some early fluctuating, achieve a degree of stability, although they plateau at a lower score than the corresponding training scores.

In conclusion, the analysis of Dice scores across the epochs indicates that models trained with augmented data exhibit improved performance. The stabilization of validation scores, in particular, points towards enhanced generalization capabilities when models are exposed to a richer variety of training examples. The data augmentation was able to achieve level of robustness that appears to mitigate overfitting.

From F1 and DICE scores results (See Figure 6a and Figure 6b), we can see data augmentation indeed helps us to improve the performance of segmentation, where two scores are overall higher in data augmentation case for all three modal types.

Model	Class	Acc.	Prec.	Rec.	Spec.	F1	Dice
Augmented_FLAIR	0:NoTumor	0.9906	0.9931	0.9972	0.8031	0.9951	0.9951
Augmented_T1CE	0:NoTumor	0.9876	0.9917	0.9955	0.7623	0.9936	0.9936
Augmented_T1CE-FLAIR	0:NoTumor	**0.9916**	0.9924	**0.9989**	0.7816	**0.9956**	**0.9956**
FLAIR	0:NoTumor	0.9887	0.9938	0.9945	0.8248	0.9941	0.9941
T1CE	0:NoTumor	0.9835	0.9868	0.9963	0.6202	0.9915	0.9915
T1CE-FLAIR	0:NoTumor	0.9895	0.9931	0.9961	0.8021	0.9946	0.9946

Table 1: Performance metrics of different models "0:NoTumor"

Model	Class	Acc.	Prec.	Rec.	Spec.	F1	Dice
Augmented_FLAIR	1:NecroticCore	0.9900	0.6010	0.2469	0.9982	0.3500	0.3500
Augmented_T1CE	1:NecroticCore	**0.9937**	0.8035	**0.5616**	0.9985	**0.6611**	**0.6611**
Augmented_T1CE-FLAIR	1:NecroticCore	0.9934	0.7912	0.5389	0.9984	0.6411	0.6411
FLAIR	1:NecroticCore	0.9894	0.5606	0.1501	0.9987	0.2368	0.2368
T1CE	1:NecroticCore	0.9917	0.8328	0.2970	0.9993	0.4379	0.4379
T1CE-FLAIR	1:NecroticCore	0.9912	0.6803	0.3690	0.9981	0.4785	0.4785

Table 2: Performance metrics of different models for class "1:NecroticCore".

Model	Class	Acc.	Prec.	Rec.	Spec.	F1	Dice
Augmented_FLAIR	2:Edema	0.9877	0.6172	0.7395	0.9920	0.6729	0.6729
Augmented_T1CE	2:Edema	0.9864	0.6051	0.5987	0.9932	0.6019	0.6019
Augmented_T1CE-FLAIR	2:Edema	**0.9912**	**0.7766**	0.6849	**0.9966**	**0.7279**	**0.7279**
FLAIR	2:Edema	0.9866	0.5792	0.8046	0.9898	0.6735	0.6735
T1CE	2:Edema	0.9847	0.5607	0.4991	0.9932	0.5281	0.5281
T1CE-FLAIR	2:Edema	0.9879	0.6327	0.6978	0.9929	0.6637	0.6637

Table 3: Performance metrics of different models for class "2:Edema".

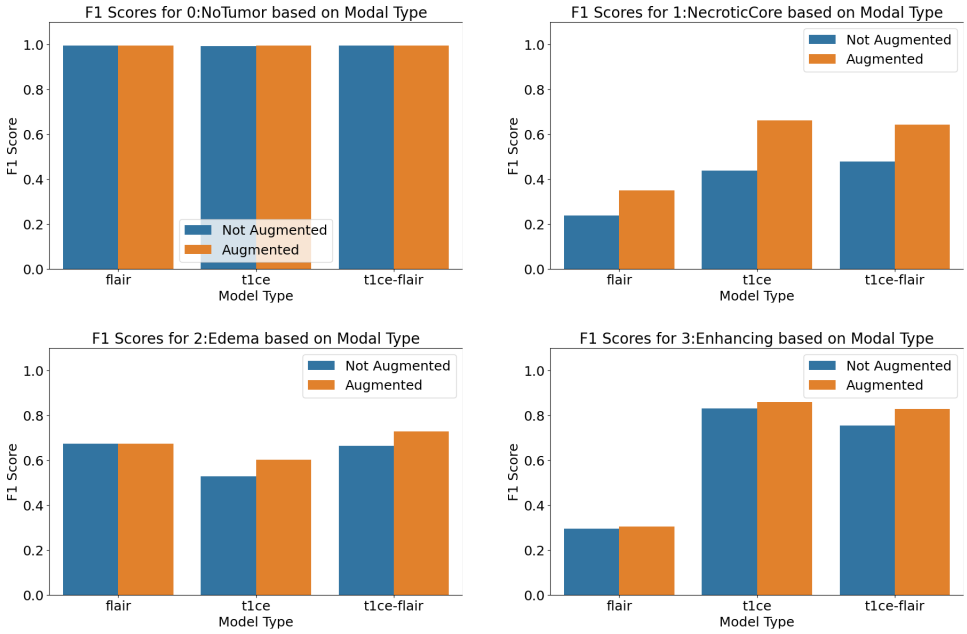
Model	Class	Acc.	Prec.	Rec.	Spec.	F1	Dice
Augmented_FLAIR	3:Enhancing	0.9924	0.3322	0.2812	0.9966	0.3046	0.3046
Augmented_T1CE	3:Enhancing	**0.9984**	0.8760	**0.8412**	0.9993	**0.8583**	**0.8583**
Augmented_T1CE-FLAIR	3:Enhancing	0.9981	0.9009	0.7667	0.9995	0.8284	0.8284
FLAIR	3:Enhancing	0.9911	0.2773	0.3121	0.9952	0.2937	0.2937
T1CE	3:Enhancing	0.9981	0.8633	0.8010	0.9992	0.8310	0.8310
T1CE-FLAIR	3:Enhancing	0.9970	0.7389	0.7715	0.9984	0.7549	0.7549

Table 4: Performance metrics of different models for class "3:Enhancing".

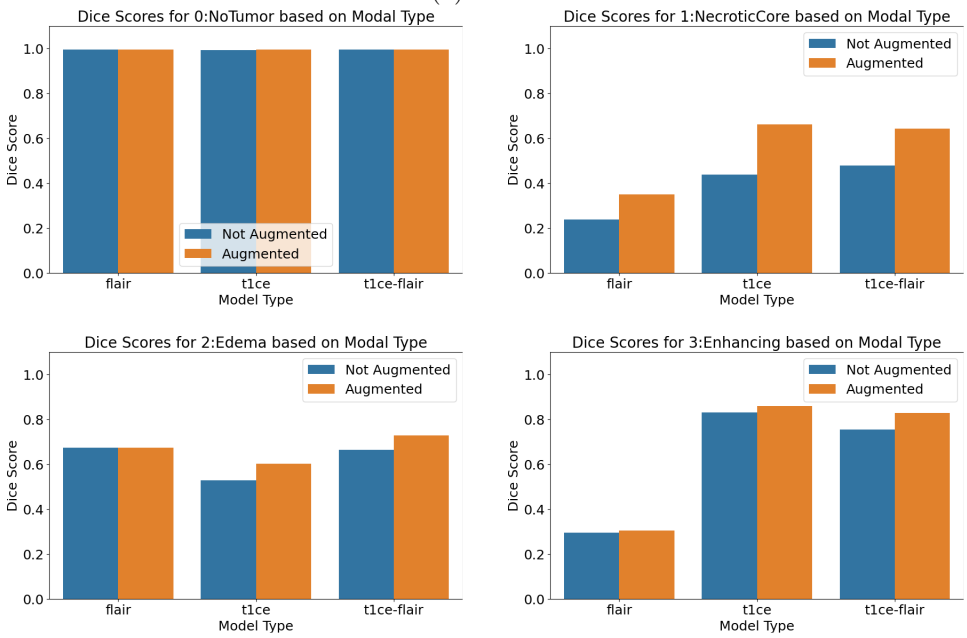
We can also see that results with T1CE modality performs much better than that with FLAIR modality in identifying Class 1 (Necrotic/Core) and Class 3 (Enhancing Tumor). At the same time, scores from FLAIR modality outperforms that from T1CE modality when classifying Class 2 (Edema).

With the booster information from FLAIR, we can see the result with T1CE-FLAIR obtains the best scores when classifying Class 0 (No Tumor) and Class 2 (Edema).

In one of the test example results (See Figure 7), we can also see T1CE's ability of capturing Necrotic Core (in coral color) and Enhancing Tumor (in gold color) and FLAIR's ability to capture Edema (in teal color). And T1CE plus FLAIR combines the advantages of both modality and capture the best result.

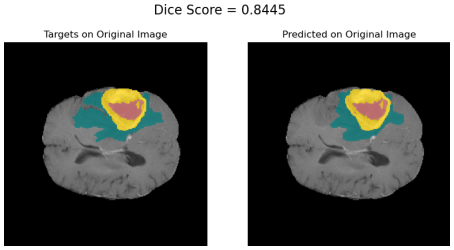


(a) F1 Scores

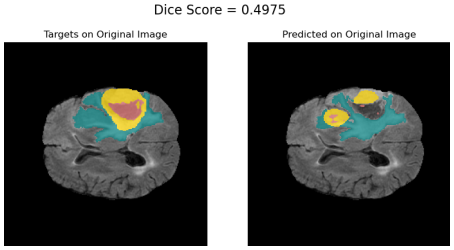


(b) Dice Scores

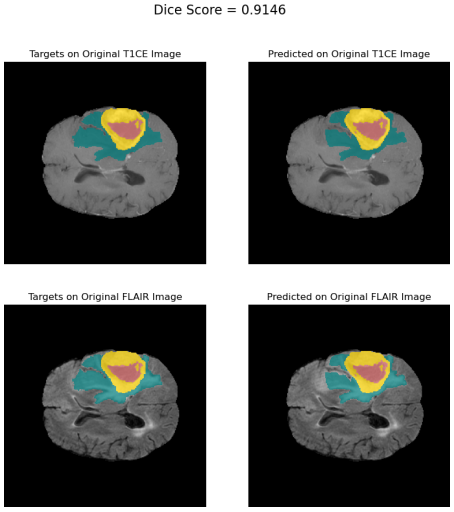
Figure 6: For each label class, scores are plotted across different model types and augmentation settings. Plot (a) displays the F1 Scores, and Plot (b) shows the Dice Scores.



(a) T1CE Example Test Result



(b) FLAIR Example Test Result



(c) T1CE-FLAIR Example Test Result

Figure 7: Example test results across different imaging modalities. Each plot illustrates the ground truth segmentation on the left and the predicted segmentation overlaid on the original input image on the right. For T1CE-FLAIR cases, predictions are displayed on both image types. Segmentation colors are as follows: coral for Necrotic Core, gold for Enhancing Tumor, and teal for Edema. Dice scores are calculated exclusively for these three tumor classes.

Conclusion

In conclusion, the aim of this project was to develop and evaluate a deep learning approach for the classification and segmentation of brain tumors, employing the classic U-Net model architecture for both single and multi-modality scenarios. Due to a limited dataset, we incorporated various data augmentation techniques. Our findings indicate that models equipped with data augmentation methods consistently outperformed those without.

Clinically, the Tumor Core (TC) comprises the primary bulk of the tumor, typically targeted for resection, including the Enhancing (Class 3) and Necrotic Core (Class 1) components. The Whole Tumor (WT) encompasses the total disease extent, combining TC (Classes 1 and 3) and Edema (Class 2). Our results show that while the augmented T1CE model demonstrated superior in segmenting Classes 1 and 3, it was less effective than the augmented FLAIR model in Class 2 segmentation. Moreover, the combination of T1CE and FLAIR modalities proved most effective in Class 0 and Class 2 segmentation.

The comparison between single-modality and multi-modality models revealed that integrating features across multiple scales and modalities significantly improve the performance of brain tumor segmentation. Specifically, the combined use of FLAIR and T1CE modalities provides a more thorough analysis by leveraging the strengths of each imaging technique. However, relying solely on T1CE information may not suffice for comprehensive tumor characterization depending on the segmentation task.

Although this project focused on the combination of T1CE and FLAIR, it is plausible that a multi-modality model incorporating additional modalities could surpass the T1CE model in all tumor class segmentation. Further research is required to confirm this hypothesis, but preliminary results suggest that merely increasing the number of modalities may not necessarily enhance model performance.

Ultimately, by achieving more precise segmentation and classification of brain tumors, this project contributes to the broader objective of improving non-invasive diagnostic techniques for brain tumors. Such advancements could aid in predicting tumor growth, facilitating treatment planning, and ultimately enhancing patient outcomes.

References

[1] F. J. Díaz-Pernas, M. Martínez-Zarzuela, M. Antón-Rodríguez, and D. González-Ortega, "A deep learning approach for brain tumor classification and segmentation using a multiscale convolutional neural network," Healthcare, vol. 9, no. 2, 2021. [Online]. Available: https://www.mdpi.com/2227-9032/9/2/153

[2] E. Elizar, M. A. Zulkifley, R. Muharar, M. H. M. Zaman, and S. M. Mustaza, "A review on multiscale-deep-learning applications," Sensors, vol. 22, p. 7384, 9 2022.

[3] A. M. Mostafa, M. Zakariah, and E. A. Aldakheel, "Brain tumor segmentation using deep learning on mri images," Diagnostics, vol. 13, no. 9, 2023. [Online]. Available: https://www.mdpi.com/2075-4418/13/9/1562

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

[5] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," 2015.

[6] S. A. Abdelaziz Ismael, A. Mohammed, and H. Hefny, "An enhanced deep learning approach for brain cancer mri images classification using residual networks," Artificial Intelligence in Medicine, vol. 102, p. 101779, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0933365719306177

[7] B. H. Menze, A. Jakab, S. Bauer, J. Kalpathy-Cramer, K. Farahani, J. Kirby, Y. Burren, N. Porz, J. Slotboom, R. Wiest, L. Lanczi, E. Gerstner, M.-A. Weber, T. Arbel, B. B. Avants, N. Ayache, P. Buendia, D. L. Collins, N. Cordier, J. J. Corso, A. Criminisi, T. Das, H. Delingette, Demiralp, C. R. Durst, M. Dojat, S. Doyle, J. Festa, F. Forbes, E. Geremia, B. Glocker, P. Golland, X. Guo, A. Hamamci, K. M. Iftekharuddin, R. Jena, N. M. John, E. Konukoglu, D. Lashkari, J. A. Mariz, R. Meier, S. Pereira, D. Precup, S. J. Price, T. R. Raviv, S. M. S. Reza, M. Ryan, D. Sarikaya, L. Schwartz, H.-C. Shin, J. Shotton, C. A. Silva, N. Sousa, N. K. Subbanna, G. Szekely, T. J. Taylor, O. M. Thomas, N. J. Tustison, G. Unal, F. Vasseur, M. Wintermark, D. H. Ye, L. Zhao, B. Zhao, D. Zikic, M. Prastawa, M. Reyes, and K. Van Leemput, "The multimodal brain tumor image segmentation benchmark (brats)," IEEE Transactions on Medical Imaging, vol. 34, no. 10, pp. 1993–2024, 2015.