Unsupervised Clustering Comparison for Brain Tumor Segmentation

Isaiah Gerena
School of Informatics, Computing, and Cyber Systems
Northern Arizona University
Flagstaff, AZ, USA
ig399@nau.edu

Matthew Gardner
School of Informatics, Computing, and Cyber Systems
Northern Arizona University
Flagstaff, AZ, USA
meg546@nau.edu

***Abstract*—**

I. Introduction

Accurate characterization of brain tumor sub regions is a critical component of clinical division making in diagnosis, treatment planning, and monitoring. Magnetic resonance imaging (MRI) provides rich multi modal information about tumor morphology, yet converting these complex patterns into structured, clinically meaningful regions is time consuming, subjective, and difficult to scale, which motivates the exploration of automated or semi automated methods that can capture tumor structure without relying on large labeled datasets.

In this project, we investigate the ability of unsupervised clustering methods for tumor subregion identification using the BraTS 2020 dataset. Our work evaluates three widely used unsupervised clustering techniques: Fuzzy C-Means (FCM), Gaussian Mixture Models (GMM), and Spectral Clustering. Each of which represent a different modeling philosophy. FCM provides soft memberships that reflect uncertainty at tissue boundaries; GMM models voxel distributions probabilistically through Gaussian components; and Spectral Clustering captures non-linear relationships using graph based similarity structures. We assess the ability of each method to recover clinically relevant tumor sub regions by comparing cluster assignments to the BraTS ground-truth segmentation masks using Dice coefficient metrics.

Across the project, we analyze the strengths and limitations of each clustering approach, the impact of preprocessing choices, and the challenges inherent to the dataset and problem at hand. Out results provide insight into when unsupervised methods can meaningfully approximate expert defined tumor boundaries and where future work such as improved feature extraction, spatial regularization, or hybrid supervised-unsupervised approaches may further enhance performance.

II. Novelty Extension

While unsupervised clustering has been applied to medical imaging, we found very few studies that directly compared Fuzzy C-Means and Gaussian Mixture Models against the more complex Spectral Clustering approach on the BraTS dataset or similar multi modal MRI collections. This gap motivated our work, which provides a unified evaluation of all three methods under a consistent experimental setup and evaluation. Another main contribution of our project is the custom preprocessing pipeline we developed to prepare the BraTS 2020 data for clustering, including modality loading, intensity normalization, mask generation, feature extraction, and voxel flattening/reshaping procedures tailored specifically for unsupervised tumor segmentation. Together, these elements create a systematic and comparable framework that highlights how each clustering technique behaves when exposed to the same inputs.

III. Background

A. MRI for Brain Tumor Segmentation

Magnetic Resonance Imaging (MRI) is the primary non invasive tool for diagnosing and monitoring gliomas due to its ability to capture derailed soft tissue contrast. Different MRI modalities highlight distinct biological properties: FLAIR suppresses cerebrospinal fluid to emphasize edema, T1 and T1CE reveal tissue structure and contrast enhancement, and T2 emphasizes fluid based abnormalities. Automated tumor segmentation seeks to convert these complementary signals into structured tumor regions that can support treatment planning, surgical navigation, and longitudinal assessment. However variations in scanner settings, patient motion, and tumor appearance make automated segmentation challenging.

B. Tumor Regions

The BraTS dataset defines three clinically relevant tumor subregions that serve as a benchmark for segmentation tasks:

- **Enhancing Tumor (ET):** Hyperintense regions on T1CE corresponding to active tumor components.
- **Tumor Core (TC):** Includes the enhancing tumor and non-enhancing necrotic or solid core.
- **Whole Tumor (WT):** Encompasses all visible tumor-related abnormalities, including edema, necrosis, and enhancing regions.

C. Unsupervised Machine Learning

Unsupervised learning aims to uncover structure within data without access to labels. In medical imaging, this is particularly appealing because manual annotation is expensive

and time consuming. For tasks such as tumor segmentation, unsupervised methods attempt to group voxels into coherent clusters based on similarity in intensity or derived features. These techniques can reveal underlying tissue categories, provide initialization for downstream supervised models, or support exploratory analysis when labeled datasets are limited.

D. Clustering Methods

Clustering algorithms differ in how they define similarity and partition data.

a) Fuzzy C-Means (FCM): FCM assigns each voxel a degree of membership to each cluster rather than a single hard label. This allows the model to capture uncertainty at tumor boundaries and is well-suited for medical images where tissues blend gradually.

b) Gaussian Mixture Models (GMM): GMM models the intensity distribution as a weighted sum of Gaussian components. Each component corresponds to a cluster, and voxel assignments are determined probabilistically. GMM can flexibly represent multi modal intensity patterns and is widely used for soft tissue classification.

c) Spectral Clustering: Spectral clustering constructs a similarity graph between voxels and uses eigenvectors of the graph Laplacian to embed data into a space where clusters are more separable. This makes it powerful for capturing non-linear relationships that intensity based models struggle with. However, it is computationally expensive and sensitive to parameter choices such as the affinity matrix and neighborhood scale.

E. Principle Component Analysis

Principle Component Analysis (PCA) is a linear dimensionality reduction technique that projects high dimensional data onto orthogonal components capturing the greatest variance. In clustering applications like this project, PCA can reduce noise, compress modalities into fewer features, and improve computation efficiency especially for high resolution 3D MRI volumes. However, PCA may also discard subtle yet clinically important variations or collapse distinctions between tumor tissues. PCA was tested as an optional preprocessing step, allowing us to evaluate whether the process benefited or distracted the clustering performance.

IV. Dataset: BraTS-2020

This study uses the Multi modal Brain Tumor Segmentation Challenge (BraTS) 2020 dataset, a widely adopted benchmark for evaluating tumor-segmentation algorithms. The dataset contains pre-operative MRI scans from patients diagnosed with gliomas, collected from multiple institutions using varying scanner models and acquisition protocols. Each subject includes four MRI modalities: T1, T1 with contrast enhancement (T1CE), T2, and FLAIR. Together providing complementary visual information about tumor structure, edema, necrosis, and enhancing regions.

The BraTS dataset provides labeled segmentation masks for three tumor sebregions:

- Enhancing Tumor (ET)
- Tumor Core (TC)
- Whole Tumor (WT)

Although ground truth labels are available, as this is a unsupervised project they are only used for evaluation.

Each subjects MRI volumes are provided as 2D NfTI files, coregistered to the same anatomical template, skull stripped, and resampled to an isotropic 1mm resolution. This standardized preprocessing ensures consistency across patients and supports voxel-wise analysis.

For this project, we used the BraTS 2020 training set, which contains 369 patients, each with four MRI modalities. Although this corresponds to 1,476 individual MRI volumes. The dataset also includes separate validation and testing sets, but these were not used in our study, as processing and managing the full training set alone presented significant computational and memory challenges.

V. Preprocessing

BraTS 2020 MRI volumes are not raw clinical scans; they are pre-skull-stripped, co-registered across modalities, and a bias field correction [1]. However, additional transformations are still required to prepare the data for unsupervised clustering algorithms such as skull stripping, feature stacking, and defining a region of interest (ROI). The following subsections detail the preprocessing steps required before execution of baseline clustering algorithms. This pipeline aims to act as a consistent source of data preparation for numerous unsupervised clustering algorithms. Below is a visualization of the workflow to prepare BraTS files for clustering implementations

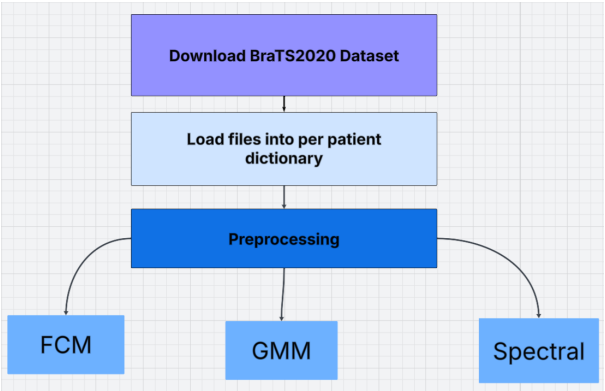


Fig. 1: Overview of the preprocessing and clustering workflow. The BraTS2020 dataset is first downloaded and loaded into a dictionary labeled `"patients"`. After preprocessing, the data branches into three clustering implementations: FCM, GMM, and Spectral Clustering.

A. Applying Brain Mask

For the purpose of this project, a binary brain mask was used to remove all non-brain voxels from MRI scans. To explain, if a voxel had an intensity greater than zero in any of the four modalities, it was considered part of the brain. This not only deals with the class imbalance of tumor and non-tumor voxels,

but also benefits clustering algorithms that utilize distance-based similarity and would form a massive cluster altering true results.

B. Intensity Clipping & Z-Score Normalization

Intensity clipping is a technique used to handle outliers in the data and avoid rare anomalies from skewing cluster representations. This technique involves clipping intensities of brain voxels to the 1st and 99th percentiles, proposed by Isensee et al. [2] in their nnU-Net framework. This step reduces the influence outliers have on model performance. This technique is provided separately to each modality, maintaining their contrast while further preparing scans for clustering.

Z-score normalization standardizes the intensity values of each MRI modality so that they share a mean and standard deviation value, which is essential because MRI intensities can vary widely across patients and modalities. This process ensures that all modalities have comparable intensity values for distance calculations, allowing the model to utilize every view of the patients brain.

C. Box Cropping & Region of Interest

To limit unnecessary computations, the smallest 3D box that contains all brain voxels is extracted and then cropped around each modality. Creating this box removes empty areas surrounding the brainmask and speeds up clustering algorithms while reducing memory usage. By defining a cropping size that includes every voxel in the brain mask the bounding box avoids removing tumor regions within the mask. Then to further aid the clustering algorithms, a region of interest is defined using non-zero values from FLAIR modalities. These values are used to define ROI because FLAIR is used to suppress cerebrospinal fluid (normal brain fluid) and improve the visibility of pathological or tumor-like fluid within the brain [3]. Therefore, Flair values of 0 are more likely to include normal brain tissue and should be avoided or shown in the background cluster if it is included. Thus, we define our ROI to assist the model when attempting to cluster heterogeneous scans, some of which contain little to no tumor regions at all.

D. Voxel-Wise Feature Stacking

Voxel-wise feature stacking consists of collecting the intensity of each voxel in the ROI of all four MRI scans and using those values to create a single row of a table. In other words, if an image had ten voxels, the table would require ten rows with four columns, each column representing a different modality. The values in the feature matrix are meant to describe different representations of brain tissue for the model to distinguish; including edema, necrotic, FLAIR enhancement, tumor core, and normal brain tissue. After stacking values, scikitlearn's StandardScaler is used to transform the data into clean and normalized voxel features that can be used effectively during clustering algorithms, and will not be dominated by a single modality.

E. Principle Component Analysis

Bishop [4] defines dimensionality reduction as a technique used to reduce noisy or redundant features in high dimensional datasets while preserving as much relevant information as possible. Principal Component Analysis (PCA) is a dimensionality reduction technique that captures the most crucial patterns in the data and reduces its dimensionality using principal components, which represent the directions of maximum variance [4]. PCA was tested across all three clustering implementations, and the results were compared when dimensionality reduction was disabled, highlighting that only spectral clustering benefited from the technique because of the memory usage of the model. GMM and FCM performance declined when enabling PCA due to the importance of all 4 modalities or dimensions being crucial to defining smooth edges during clustering. Dimensionality reduction is a technique commonly used in unsupervised machine learning, and although the results declined, its implementation was crucial for educational purposes.

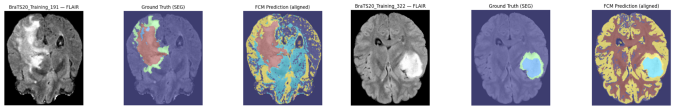
VI. Clustering Implementations

A. Fuzzy C-Means

A Fuzzy C-Means implementation was applied to the scaled voxel feature matrix, as previously stated PCA was disabled for this model after comparison of Dice results. After execution of the preprocessing pipeline, each voxel inside the ROI was represented as a four dimensional vector consisting of each modality. Using the scikit-fuzzy implementation of the C-Means algorithm, the feature vector was passed into the clustering algorithm to produce 3 clusters and attempting to completely exclude healthy brain tumors instead of clustering them like GMM and Spectral implementations. The model was configured with a fuzzifier parameter m = 2.0 m=2.0, a maximum of 300 iterations, and a convergence tolerance of 1e-5 allowing the algorithm to gradually improve memberships and cluster centroids.

After convergence, voxels were assigned to the cluster corresponding to its highest membership score in an attempt to define smooth segments of brain tumor regions. Resulting clusters were reshaped into the 3D ROI and aligned using a BraTS specific ground truth label for visual analysis and later explained metrics. Clusters were mapped using three different labels; Whole tumor (WT), Enhancing Tumor (ET), and Tumor Core (TC) for comparable visualization to the true segmentation files and evaluation of dice, IoU, precision, and recall scores.

As seen in previous research, FCM has a tendency to cluster the entire brain when applied to brain tumor segmentation because base FCM tends to treat large healthy portions of the brain as part of clusters [5]. As illustrated in **Figure 2**, when compared to FLAIR and true segmentation files FCM fails to capture distinct boundaries of tumor subregions and instead collapses to random intensity noise by clustering the entire brain. This analysis was validated during evaluation with the presence of a 1.00 Whole tumor recall score.



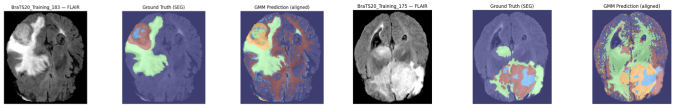
(a) FCM Result Example 1 (b) FCM Result Example 2

Fig. 2: **Figure 2** shows two representative visualizations for the FCM clustering results. Each visualization displays three aligned slices: **Left** – FLAIR modality, **Center** – Ground truth segmentation, **Right** – FCM cluster assignment.

B. Gaussian Mixture Model

The Gaussian Mixture Model (GMM) was implemented on a per patient basis using the voxel wise feature matrix derived from the region of interest (ROI). For every voxel inside the ROI, the four modality intensities were stacked into a four dimensional feature vector, standardized using the scikit-learns Standard Scaler, and reduced to three components using PCA. These PCA features were than passed into scikits learns Gaussian Mixture algorithm with four mixture components, diagonal covariance matrices, a single initialization run, a maximum of 50 iterations, and a fixed random seed for reproducibility.

Once the GMM has been fit, each voxel was assigned to the mixture component with the highest posterior probability, producing a 1D array of cluster labels. These labels were reshaped back into the cropped 3D brain volume using the ROI mask. The GMM outputs were aligned to the BraTS segmentation labels using a Hungarian matching procedure. For each patient a confusion matrix was computed between the cluster assignments and the ground truth labels, and the optimal mapping was applied to relabel the clusters. These aligned clusters were than used to compute WT, TC, and ET masks, which were evaluated with Dice, precision recall, and IoU metrics.



(a) GMM Result Example 1 (b) GMM Result Example 2

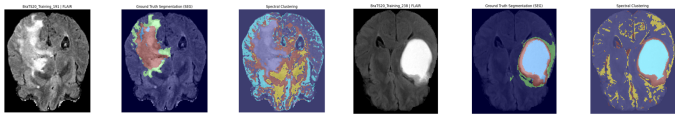
Fig. 3: **Figure 2** shows two representative visualizations for the GMM clustering results. Each visualization displays three aligned slices: **Left** – FLAIR modality, **Center** – Ground truth segmentation, **Right** – GMM cluster assignment.

Dice scores suggest that GMM outperformed spectral clustering however, visual analysis proves the spike in GMM's dice was due to random speckles luckily behind placed inside the smallest region. Similar to other clustering implementations, GMM assignments overfit to the ROI often resulting in high recall with poor precision scores. The consistent overlap of the edema and healthy brain tissue highlights the models ability to identify actual tumors, but underscores its failure to recognize abnormalities in the brain caused by tumors.

C. Spectral Clustering

Spectral Clustering was implemented using a sub sampling and label propagation strategy to make the method computationally feasible for the large 3D MRI volumes on the compute hardware available. As with GMM, voxel level intensities from all four modalities were standardized and reduced to three PCS components. Because full spectral decomposition on hundreds of thousands of voxels is far too large to use, each patients ROI was randomly sub sampled to a maximum of 75000 voxels. These selected voxels were used to build a k nearest neighbors affinity graph with 40 neighbors. Spectral clustering was then performed on this subset using four clusters, a nearest neighbor affinity matrix, ARPACK as the eigen solved, a k means for label assignment.

After clustering the sub sampled points, their labels were propagated to all remaining voxels using a k nearest neighbor classifier with five neighbors. This produced full resolution spectral labels for every voxel inside the ROI, which were than send back into the cropped 3D volume. Similar to the GMM pipeline, a Hungarian alignment step was applied to match the arbitrary cluster IDs to the BraTS tumor classes by maximizing agreement between predicted clusters and ground truth labels. The aligned spectral labels were used to generate WT, TC, and ET masks and obtain the performance metrics shown in the results section and below.



(a) Spectral Result Example 1 (b) Spectral Result Example 2

Fig. 4: **Figure 2** shows two representative visualizations for the Spectral clustering results. Each visualization displays three aligned slices: **Left** – FLAIR modality, **Center** – Ground truth segmentation, **Right** – Spectral cluster assignment.

As illustrated in **Figure 4b**, Spectral Clustering produced results that were visually smoother and more spatially structured than both GMM and FCM, despite its comparatively low dice scores. It is important to note that this method was limited to 75,000 from an ROI containing nearly 1,000,000 voxels per patient, highlighting the strengths and limitations of this reduced implementation. Visualization suggests that the algorithm is highly sensitive to peripheral noise, proven by a distinct outline-like cluster forming along the edges of the brain. **Figure 4a** suggests the algorithm suffered from similar limitations as FCM, with the entire tumor in a single cluster and the rest filling almost the entire brain.

VII. Performance Evaluation

Before moving to evaluation, it is important to note that the BraTS2020 dataset included numerous files without actual tumors present in the FLAIR or true segmentation files, resulting in under-represented metrics for all three clustering methods. This is not a limitation of the algorithm but of

the dataset because the purpose of this implementation is tumor segmentation that differs from tumor detection. Future segmentation should discard healthy brain files and only retain those that include some type of tumor region. To continue, clustering methods were evaluated on metrics including; dice score, precision, recall, and Intersection over Union (IoU).

Dice score is used to measure the overlap between clustered segments and the ground truth segmentation files, emphasizing the models ability to capture distinct boundaries between each tumor region. Lower Dice scores reflect missed regions, excessive false positives, or the presence of fragmentation or speckled noise in random assignments around the brain. In contrast, higher Dice scores indicate strong spatial agreement with the ground truth segmentation files, demonstrating the model's ability to capture smooth and coherent boundaries between tumor subregions.

Precision evaluates the amount of predicted voxels that are actually correct, making it sensitive to false positives in over-clustered predictions. A low precision reflects that the model continuously labels healthy tissue as tumor. Recall monitors the amount of tumor voxels that the model successfully captures, showing if the model can successfully learn properties of tumor tissue. However, a higher recall can also be an indication of overfitting by classifying a majority of voxels as tumors to avoid misses.

IoU, similar to dice, compares how much the predicted tumor region overlaps with the ground truth but penalizes false positives and false negatives more strictly than dice scores. This metric often results in lower results than dice, but provides insight into how well the model capture the true tumor regions.

A. Fuzzy C-Means

FCM resulted in poor dice scores for tumor segmentation, as expected compared to prior research [5]. The habit of consistently clustering the entire brain is the main factor for resulting dice scores and highlights the entire tumor often belonging to one cluster seen during visual analysis. The use of more spatially aware FCM implementations is needed before proceeding stacking FCM as an unsupervised component in hybrid algorithms.

Precision and recall scores mimic behavior seen during visual analysis, high recall is reflected by clustering expanding the entire brain with the inability to distinguish between healthy and malicious regions shown in precision metrics.

Similar to dice, IoU for FCM reflects the models inability to capture the whole tumor region and show an ability to create any type of subregion within the tumor. Overall, FCM failed to segments the varying parts of brain tumor and highlight the need for more spatially aware implementations of a FCM concept.

TABLE I: FCM Mean Per-Class Segmentation Metrics

Region	Dice	Precision	Recall	IoU
WT	0.1934	0.1116	1.0000	0.1116
TC	0.0806	0.0465	0.5041	0.0456
ET	0.0610	0.0429	0.3565	0.0392

B. Gaussian Mixture Models

Guassian Mixture Models achieved the strongest Dice and IoU results out of all three clusters, although spectral clustering was reduced to limited voxels due to computing limitations. Scores reflect GMMs ability to capture the internal tumor core while failing to separate swelling caused by tumors from healthy brain tissue. Although GMM shows a significant increase in performance when compared to other models, its scores are inflated due to fragmentation accidentally landing in the smallest region of the Non-enhancing tumor core.

Precision and recall show that the model rapidly learned to recognize tumor voxels but failed to create smooth edges between tumor edges and healthy brain tissue. The model consistently over-clustered the edema region of the brain resulting in impressive recall scores for the whole tumor label. However, Tumor core recall shows the model could slightly differentiate between internal tumor regions.

TABLE II: GMM Mean Per-Class Segmentation Metrics

Region	Dice	Precision	Recall	IoU
WT	0.3115	0.2179	0.8055	0.1992
TC	0.3064	0.2324	0.7188	0.2011
ET	0.2675	0.3013	0.3956	0.1980

C. Spectral Clustering

As mentioned previously, spectral clustering capabilities were reduced for resource efficient purposes and resulted in under performing scores even with significant training time. Dice and IoU scores show that the model failed to capture any significant adequate representation of tumor regions, however, visualization suggests that without choke-holding the model it may have outperformed GMM. Although dice and IoU are more similar to FCM clustering, visualization shows that this method was able to learn internal tumor regions but plateaued due to its habit of clustering edges of the brain. The training time required for calculations was not reflected in results, making GMM more attractive due to limited training time and high scores.

Precision and recall once again reflect the inability to capture the smooth edges of segmentation anywhere near the level of the ground truth file. These metrics validate that bare clustering would be futile for deployable brain tumor segmentation due to drastically different representations of brain tumors that the model can not consistently capture. Hybrid or enhanced version of clustering methods may yield more optimal results.

TABLE III: Spectral Clustering Mean Per-Class Segmentation Metrics

Region	Dice	Precision	Recall	IoU
WT	0.1541	0.1027	0.6282	0.0894
TC	0.1184	0.0713	0.4872	0.0689
ET	0.0637	0.0395	0.2107	0.0379

VIII. Discussion & Analysis

Poor results achieved by clustering algorithms are a combination of a dynamic dataset and the inherent limitations when

attempting to cluster MRI type data. However, visual analysis suggests these models can serve as methods for initialization or the configuration of pseudo-labels for more advanced deep learning techniques. Methods like GMM and FCM present that capability to identify tumor regions, but fail when being evaluated on specific segmentation metrics. Consistently high recall scores across models suggest a habit of overfitting and an inability to distinguish between normal brain tissue and mainly the edema subregion. This research served as a starting point for unsupervised methods based on unbiased and public MRI data. The implemented versions evaluated were not meant to stack onto deep learning models, but to serve as a starting point for that purpose.

Future research should explore the possibility of a spatial FCM implementation along with more optimized versions of GMM due to its speed and initial results. Incorporating spatial awareness or neighborhood aggregation could reduce the fragmentation caused by random noise and improve the precision of tumor subregions. Expanding the dataset to exclude healthy brain images would further enhance its applicability to this research, as the lowest scores usually consisted of files without a segmented tumor region. These suggestions provide a foundation to transform simple clustering into segmentation of complex brain tumors.

IX. Conclusion

This project presented a unified comparison of the three unsupervised clustering methods; Fuzzy C-Means, Gaussian Mixture Models, and Spectral Clustering on brain tumor segmentation using the BraTS 2020 dataset. By developing a consistent preprocessing pipeline and applying voxel level features across al models, we were able to evaluate each methods behavior under comparable conditions. The quantitative and qualitative results show clear differences in how these clustering techniques interpret multi modal MRI data, with GMM achieving the most highest performance across tumor subregion, FCM suffering from severe overfitting to the full brain mask, and Spectral Clustering demonstrating potential but struggling with fragmentation and computational constraints.

Across al models, the overall performance highlights the difficulty of using these clustering methods alone to recover the complex biological tumor regions. Nonetheless, this study provides valuable insight into where these techniques succeed, where they fail, and what types of preprocessing and feature representation most strongly influence the outcome.

While these methods are not replacements for modern supervised segmentation methods, they remain a useful framework for exploratory analysis, dataset characterization, and initialization of more advanced models. Future work could integrate spatial regularization, incorporate multi model feature engineering beyond intensity values, or explore hybrid semi supervised approaches that leverage a small number of labeled examples. These directions may help bridge the gap between traditional clustering algorithms and the clinically meaningful segmentation performance required for real world applications.

References

[1] B. H. Menze et al., "The multimodal brain tumor image segmentation benchmark (brats)," *IEEE Transactions on Medical Imaging*, vol. 34, no. 10, pp. 1993–2024, 2014. DOI: 10.1109/TMI.2014.2377694.

[2] F. Isensee, P. F. Jaeger, S. A. A. Kohl, J. Petersen, and K. H. Maier-Hein, "Nnu-net: A self-adapting framework for u-net-based medical image segmentation," *Nature Methods*, vol. 18, no. 2, pp. 203–211, 2021. DOI: 10.1038/s41592-020-01008-z.

[3] B. H. Menze et al., "The multimodal brain tumor image segmentation benchmark (brats)," *IEEE Transactions on Medical Imaging*, vol. 34, no. 10, pp. 1993–2024, 2014. DOI: 10.1109/TMI.2014.2377694.

[4] C. M. Bishop, *Pattern Recognition and Machine Learning*. New York: Springer, 2006, ISBN: 978-0-387-31073-2.

[5] L. Hua, Y. Gu, X. Gu, J. Xue, and T. Ni, "A novel brain mri image segmentation method using an improved multi-view fuzzy c-means clustering algorithm," *Frontiers in Neuroscience*, vol. 15, p. 662 674, 2021. DOI: 10.3389/fnins.2021.662674. [Online]. Available: https://doi.org/10.3389/fnins.2021.662674.