A Comparative Study of Machine Learning and Deep Learning Multi-Class Segmentation Methods

Final Report for CS 534, Emory **GitHub Repo: https://github.com/eliznemeti/Brain_Tumor_Segmentation_Project**

ELIZABETH NEMETI and SHANIAH REECE

Abstract. Automated brain tumor segmentation which utilizes machine learning and deep learning techniques is a major research focus for enhancing the diagnostic efficiency of medical imaging. Magnetic Resonance Imaging (MRI) offers detailed images essential for tumor detection and diagnosis, yet manual tumor detection remains labor-intensive and is error-prone, especially for tumor types with complex tumor morphologies and early-stage lesions. In this paper, we highlight the challenges and advancements in automated brain tumor segmentation, through a comparative analysis of the deep learning architecture, U-NET against simpler machine learning models Support Vector Machines (SVM) and Fuzzy C-Means (FCM). Our study aims to improve model interpretability and performance metrics through a comprehensive evaluation across the diverse BRaTs 2018 and SWARTAJ datasets. Furthermore, we address the scarcity and quality of annotated data by implementing data augmentation techniques and exploring both supervised and unsupervised segmentation approaches. By refining automated brain tumor segmentation methods, our goal is to contribute to more accurate and efficient clinical decision-making in brain tumor detection.

ACM Reference Format:
Elizabeth Nemeti and Shaniah Reece. 2024. A Comparative Study of Machine Learning and Deep Learning Multi-Class Segmentation Methods [0.5em] Final Report for CS 534, Emory GitHub Repo: https://github.com/eliznemeti/Brain_Tumor_Segmentation_Project . 1, 1 (May 2024), 12 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 Introduction

Magnetic Resonance Imaging (MRI) is a non-invasive imaging technique [1, 12, 15]. Its depth and detail make it the most common and well-established imaging technique in brain tumor diagnosis. Despite this benefit, manual annotation or "segmentation" of tumor regions by radiologists is a laborious and time-consuming task [13]. Due to the sheer volume of the image sets, unclear morphology of brain tumor regions, and small size of early-stage tumors, manual segmentation can produce inaccurate results [6]. Additionally, for a patient's further treatment options, quantification of the tumor area is needed [10]. Hence, automated techniques are sought after for accurate diagnosis.

Automatic segmentation computationally categorizes pixels to distinguish between brain tissue, image background, and tumors [9]. It can be accomplished using machine learning and deep learning methods to aid in the fast and accurate detection of these tumors. These methods have the advantage of exploiting large datasets to learn the patterns and characteristics indicative of tumors, making them more feasible than manual approaches [8]. While they can exhibit high accuracy, they are still liable to generating inaccurate conclusions due to issues with generalization across diverse tumor types [7], variability in imaging protocols, and the requirement for annotated data. Thus, refinement and evaluation are necessary for effective segmentation techniques [8].

Deep learning models often outperform machine learning models in brain tumor segmentation. however, these results are not enough to trivialize traditional machine learning models [5]. For example, while research shows that deep learning models can be effective for complex image segmentation tasks, there are still instances when they are less suitable, making machine learning models more suitable options. These include working with small datasets, enhancing interpretability and computational efficiency, and handling noise and outliers [5].

1.1 Approach and Rationale

By building on recent advancements [8], we compare and evaluate deep learning models against traditional machine learning algorithms (SVM) and clustering techniques (FCM) across various performance metrics (see 5). Through a comparative analysis, we highlight opportunites and limitations of each technique for enhancing interpretability, generalizability and performance, thereby advancing the efforts to enhance automated brain tumor segmentation.

Authors' Contact Information: Elizabeth Nemeti; Shaniah Reece.

Manuscript submitted to ACM

1.2 Novelty and Justification

Profiling the state-of-the-art for segmentation reveals a strong shift towards deep learning methods, as evidenced by the significant increase in publication rates illustrated in (see Figure 7 in Appendix) over the last few years. While deep learning methods consistently outperform machine learning methods, they are hindered by the requirement for large annotated datasets, interpretability issues, and substantial computational resources. In a 2021 study utilizing our selected dataset [2] (see also Section 4.1), Maas et al. developed a U-CNN, an automatic multiclass segmentation tool [8]. Although their model yielded promising results, they identified the following areas for future work: 1) conducting a comparison of the network model to simpler architectures to enhance understanding of the underlying information extracted from medical images, and 2) employing datasets covering a wider variety of samples and instances of rare occurrences, as well as more diversified cohorts. Our research builds upon their foundations by addressing these identified opportunities that have informed our aims, further described below.

1.3 Key Contributions and Significance

The following aims are designed to address and expand on the areas of future work outlined by Maas et al. [8] and Zhou et al. [16], ensuring that our research contributes effectively to filling these identified gaps.

Aim 1: To address the need for comparing CNNs with simpler models to enhance interpretability, we will:
Compare a U-NET (CNN) with both an SVM and FCM across multiple performance metrics and qualitative visual results.
Aim 2: To address the need for utilizing more varied/diversified datasets in training, we will:
Implement data augmentation for U-NET and benchmark all three models with the well-established BRaTs 2018 dataset.
Aim 3: To address the labeling challenge in medical imaging, where labels are often sparse, noisy, or inconsistent [16], we will:
Assess the effectiveness of supervised (i.e. UNET and SVM) versus unsupervised segmentation methods (i.e. FCM).

By addressing these aims, we hope to advance the state-of-the-art in automated brain tumor segmentation, paving the way for more accurate and precise diagnostic tools. Furthermore, our work holds implications for improving patient outcomes by enhancing diagnostic precision through the advancement of state-of-the-art techniques. In the subsequent sections we provide by reviewing existing work in the field, detail our methods and model selection, outline our evaluation plan and findings, and discuss the implication of our results on the field.

2 Background

Segmentation in medical image processing can be employed using manual, semi-automatic, or fully automatic approaches. However, automation has become a necessity as radiologists require efficient techniques for accurate diagnosis using large data sets [6]. Constant improvement of medical image segmentation remains at the forefront [9]. In recent years, applications of DL approaches have been significantly greater than ML techniques [7]. We choose to compare an SVM and CNN due to their state-of-the-art nature and strong prominence in current publications that explore our same task of brain tumor segmentation. Ranjbarzadeh et al.'s 2019 comprehensive review [13] indicated SVM and CNN as being the most widely used ML and DL tools for brain tumor segmentation as of late 2022 (see supporting figure in Appendix).Their popularity stems from excellent performance; for example, SVM indicated high accuracy rates over 90 percent of the time. Although FCM was only the fifth most popular approach, we also include it in our research due to its implementation of fuzzy membership, making it particularly useful for pixel classification and hence image segmentation [9]. Including FCM also allows us to compare an unsupervised approach with our two supervised approaches, SVM and CNN. However, as FCM is only a clustering method, to obtain a segmentation outline, we will incorporate an Active Contour (AC) algorithm. To select our CNN, we opted for "one of the key contributions that emerged from the medical imaging community, the U-NET architecture" [6], developed specifically for biomedical image segmentation and proven to robustly execute segmentation tasks [14].

3 Methods

3.1 Algorithm Description, Notation, and Pertinence

In this section, we describe the algorithms used in our study, highlighting their relevance and suitability for our goal of comparing deep learning and traditional machine learning methods in MRI brain tumor segmentation.

3.1.1 Convolution Neural Network (CNN) with U-Net Architecture Our primary supervised algorithm is a CNN with U-Net architecture illustrated in Figure 1. The U-Net framework consists of an encoder-decoder pathway designed for semantic segmentation tasks. This makes

it very effective at localizing intricate structures such as the brain tumors we will be analyzing from MRI images. The encoder comprises of a series of convolutional operations (hence the name CNN) C_i, each followed by rectified linear unit (ReLU) activations and max-pooling operations:

$$C_{out} = \text{MaxPool}(\text{ReLU}(C_{in} * K + b))$$

where $*$ indicates the convolution operation, K represents the kernel, and b is the bias. The decoder up-samples the feature maps and applies convolutions to generate precise segmentation outputs, leveraging skip connections that concatenates encoder and decoder layers, for enhanced localization.

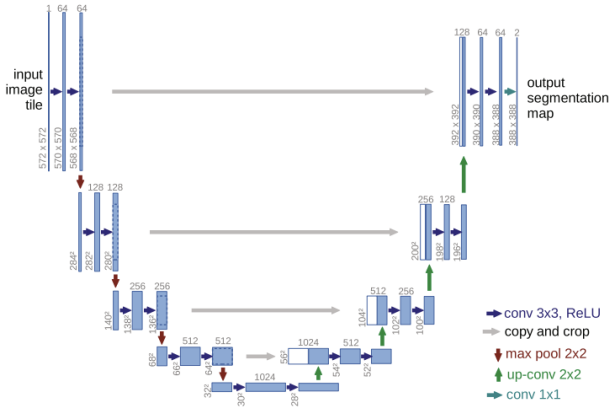

Fig. 1. The network structure of U-net from Feng and Tang 2024 [4]

3.1.2 Support Vector Machine (SVM) with Radial Basis Function (RBF) Kernel We utilize the SVM with an RBF kernel because of its effectiveness in handling high-dimensional data and capturing non-linear relationships between features. The SVM decision function is defined by:

$$f(x) = \sum_{i=1}^{n} \alpha_i y_i K(x_i, x) + b$$

where α_i, y_i, and b are the model parameters, K is the RBF kernel, x represents a data point, and x_i are the support vectors. This model is well suited for our tumor classification task due its ability delineate complex boundaries between tumor and non-tumor regions.

3.1.3 Fuzzy C-Means (FCM) Clustering The unsupervised FCM algorithm is our vehicle to explore data clustering based on similarity, a particularly useful tool when annotations are limited or ambiguous in MRI images. FCM minimizes the objective function:

$$J(U, V) = \sum_{i=1}^{m} \sum_{j=1}^{c} u_{ij}^{p} \|x_i - v_j\|^2$$

where U represents the membership matrix, V, the cluster centers, u_{ij} the degree of membership of x_i in the cluster j, m the number of data points, c is the number of clusters, and p the fuzziness index. The iterative optimization process of FCM, updating U and V, aims to partition to better segment the MRI images into meaningful clusters, in our cause potentially highlighting subtle variations in tumor presence.

Figure 2 provides an overview of our pipleine structure for each algorithm.

4 Experiments

4.1 Data Description: Brain Tumor Classification MRI Images

We have selected a comprehensive brain tumor dataset acquired from Nanfang Hospital in Guangzhou and General Hospital at Tianjin Medical University, China, between 2005 and 2010. Patient records and information are anonymized and de-identified. This dataset comprises 3064 T1-weighted contrast-enhanced MRI slices from 233 patients, covering three distinct tumor types: 708 meningiomas, 1426 gliomas, and 930 pituitary tumors. We utilized the complete dataset in our study. The images exhibit an in-plane resolution of 512×512 with a pixel size of 0.49×0.49 mm², a slice thickness of 6 mm, and a slice gap of 1mm. Each tumor border was meticulously delineated by three radiologists

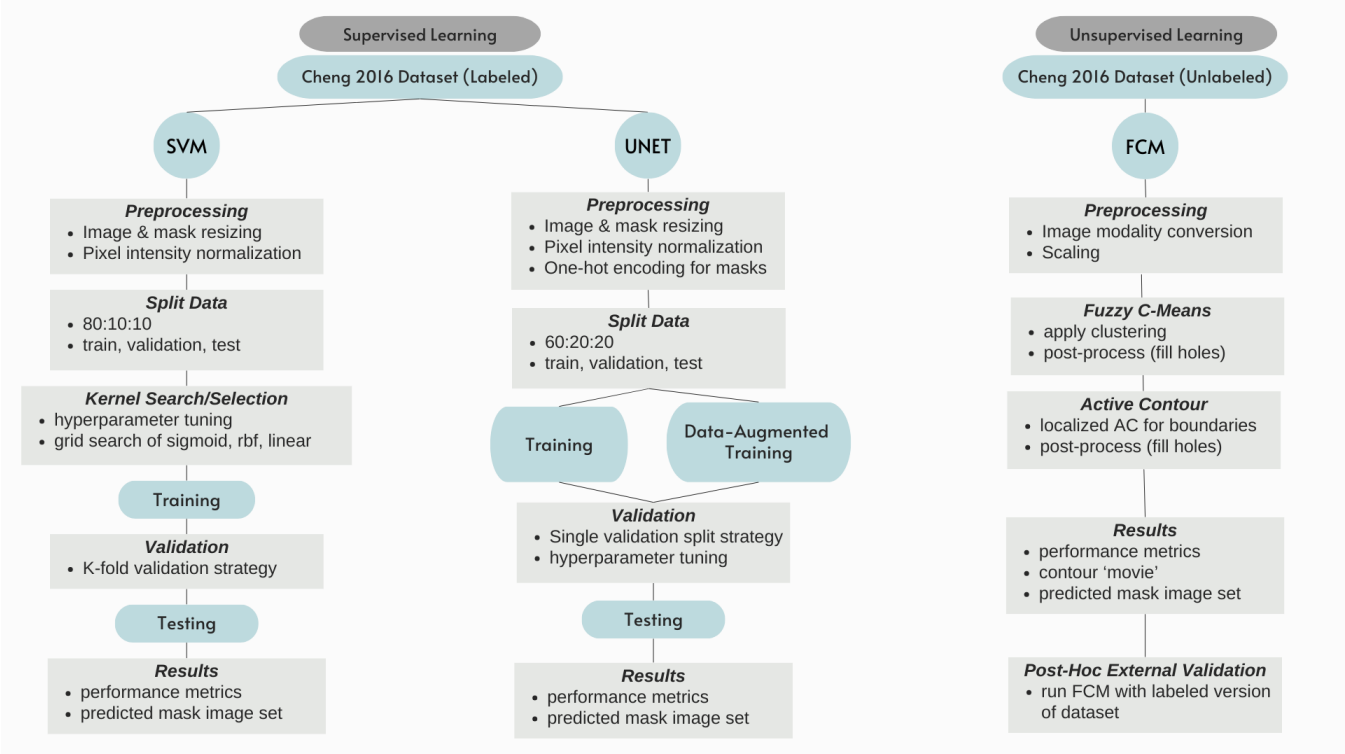

Fig. 2. Illustration of our high-level ML/DL pipeline (our figure)

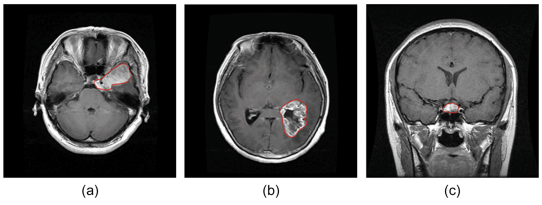

Fig. 3. Illustrations of three typical brain tumors: (a) meningioma; (b) glioma; and (c) pituitary tumor. Red lines indicate the tumor border. [3]

forming the "ground truth annotation masks" or, "labels" crucial for our segmentation tasks. Organized in MATLAB's .mat format, each file contains structured data with several fields such as tumor type, patient ID, image data, tumor borders, and binary tumor areas. Figure 3 illustrates examples of each tumor type with their corresponding ground truth annotations, highlighting the dataset's suitability for robust analysis and segmentation tasks.

4.1.1 Benchmarking Dataset Description The benchmarking dataset was sourced from: (1) BraTS 2018 (link) and (2) Kaggle Brain Tumor Classification Dataset (link). Glioma class sourced from BraTS 2018, meningioma and pituitary classes sourced from Kaggle Dataset.

The **BR**ain **T**umor **S**egmentation Challenge is a commonly used and trusted data source that contains clinically acquired 3D multimodal MRI scans, including expert-revised ground truth labels. The scans were collected from various clinical protocols and scanners from multiple institutions. From the 2018 dataset we extracted MRI of lower-grade glioma to be our benchmark glioma class. The dataset was presplit, therefore we sourced our class from the training set as it alone contained labeled data necessary to benchmark our model with. The MRI were NIfTi format (.nii.gz) across the modalities: T1, T1Gd, T2, and FLAIR, from which we extracted only T1 (no contrast enhancement (CE)). The labels were already derived by manual segmentation from up to four annotators, and were additionally approved by expert neuro-radiologists. The data includes the preprocessing steps: coregistration to same anatomical plate, interpolation to same resolution 1 mm^3, and skull stripping which we did not extrapolate to Cheng 2016 or the Kaggle dataset due to lack of access to domain experts. The degree of effort attributed to developing this dataset makes its popularity well founded and a justified selection for our study. Its disadvantage is lacking a meningioma or pituitary class. Hence, upon finding no complementary BraTS dataset to complete our three classes, we selected

a Kaggle dataset that contained meningioma and pituitary classes. This presplit, cleaned dataset consists of 3264 .jpg files from which we extracted the labeled training data. The dataset was defined by four categorical labels: glioma, meningioma, pituitary and no tumor, from which we extracted the meningioma and pituitary data. Different to BRaTS 2018, this data is T1-CE and augmented.

4.2 Exploratory data analysis, preprocessing, feature extraction, and feature selection

4.2.1 Preprocessing images and masks. We systematically loaded and preprocessed the training and benchmarking MRI images and tumor masks from the .mat files using libraries hdf5storage and cv2. **Images:** We resized images to uniform 256x256 pixel resolution and normalized them to a [0,1] scale (black [0] to white [1] pixels), to ensure consistency across all data inputs. 256x256 resolution was instated post-training following memory pressure on the original 512x512 resolution being too high for our processor. **Masks:** To prepare multi-class segmentation, we implemented one-hot encoding on tumor masks, converting them into 3-channel binary masks representing the three tumor classes: meningioma, glioma, and pituitary tumor. For the FCM algorithm, the images were stripped of their corresponding labels with an in-house script (see GIT) to execute clustering. Our preprocessing (and entire pipeline), ensures data integrity via sanity checks that do not tolerate passing to another stage upon incorrect data type and distribution, setting a robust foundation for later model training.

4.2.2 Exploratory Data Analysis and Preprocessing Plan. **Cheng 2016:** During data exploration we discovered a class imbalance within the Cheng 2016 dataset [x708 meningioma class:x426 glioma class:x930 pituitary class]. To address this imbalance, we implemented an in-house class balancing algorithm (see GIT) to ensure an equal distribution of samples across all tumor classes, preserving the maximum amount of data. **BRaTS 2018:** Since the data in our BRaTS dataset were all in .nii.gz format, we decompressed them to extract the NIFTIs, and then converted the .mat files (see GIT). We then preprocessed them to match our Cheng dataset by resizing all images to a uniform 256 x 256 pixel resolution and normalized pixel values to a scale of [0,1]. The masks corresponding to tumor annotations were encoded using a one-hot encoding scheme to represent the three tumor classes for segmentation tasks. **Sartaj2017:** Since our original Cheng dataset was produced in 2017, the formatting of the matlab files adhered to a much earlier version. As such simply converting our Sartaj dataset to .mat files was not sufficient to ensure consistency. We needed to meticulously analyze the structuring of the Cheng dataset, including its variables. This meant implementing yet another in-house algorithm to ensure compatibility between the two . We converted the images in our Sartaj dataset to Matlab v7.3 files and generated the same 1x1 struct holding the relevant variables for each image. With this change, we were successfully able to integrate this benchmark dataset into our pipeline. By standardizing preprocessing steps across our models and confirming the uniformity of data formats, we ensured that our datasets are prepared for ML/DL applications, and our pipelines are well reproducibile for subsequent analyses.

4.2.3 MRI Feature Extraction and Selection. **UNET**'s depth comes from its many layers, but the takeaway is they're actually composed of filters (see again Fig.3) that are responsible for **feature extraction** during segmentation. The contracting path (encoder layers) best captures the contextual information from the image (high-resolution features) while reducing its spatial information. Maxpooling helps enhance that process. However, to make the segmentation map, the spatial information needs to now be combined with the feature information up the expansive pathway (decoder layers) to allow for accurate localization. This is where upsampling can enhance this pathway. UNET's deep learning architecture bypasses a need for manual **feature selection** since it can automatically determine the best features for segmentation. It does so by preserving essential features via max pooling and up-sampling meanwhile allowing lesser features to wane out of focus. **FCM+AC** approaches feature extraction and selection entirely different. The FCM algorithm extracts features by assigning each pixel a probability of belonging to different clusters based on their intensity values, allowing for the capture of subtle intensity variations within the MRI images. This soft clustering is particularly adept at handling the inherent ambiguities in medical images. To achieve feature selection, the Active Contour algorithm adjusts the clusters to conform to the ground truth tumor boundaries. The conformation occurs by minimizing a function that balances image fidelity against contour smoothness. This in turn, ensures that the more relevant features for the segmentation task are retained. Finally, SVM supports feature selection and extraction by highlighting relevant features through decision boundary optimization. Unlike explicit selection methods such as FCM, SVM uses the kernel trick to project data into higher dimensional spaces, revealing more complex relationships among features. Parameters such as kernel choice and the regularization parameter, can be tuned to influence this mapping, allowing for a closer analysis of important feature behavior. The algorithm then finds the optimal hyperplane for all the features in the data and assigns features closer to the hyperplane to their respective class. This characteristic of SVM makes it a flexible and suitable approach for feature-centric tasks like brain tumor classification.

4.3 Modeling Choices: Model Selection, Determining Parameters, Design Choices

4.3.1 Defining UNET. The UNET model was selected due to its optimization for biomedical image segmentation and its state of the art nature. In addition to the general architecture discussed in Methods (see Section 3), we specifically designed the UNET to optimize performance for MRI brain tumor segmentation. The resolution parameter is increased to 512 and decreased to 128 during tuning. Each convolutional layer in both encoder and decoder pathways includes kernel initializers to maintain the stability of weights during training. We additionally, follow encoding with batch normalization to ensure training stability as well.A dropout rate of 0.3 is applied post max-pooling in the second encoder block and before the last upsampling in the decoder to mitigate overfitting. The output layer employs a softmax activation function to classify each pixel into one of three tumor types. Training involves 30 epochs (adjusted in tuning), a batch size of 4 (adjusted in tuning), to ensure sufficient gradient updates per epoch. The Adam optimizer is used for its adaptive learning rate, reducing the need for manual tuning. The loss function is a hybrid that combines categorical cross-entropy for accuracy and dice loss to address class imbalance by maximizing overlap between the predicted and true masks. Additional metrics such as multi-class specificity, accuracy, recall, and precision provide comprehensive performance evaluation.

4.3.2 Defining FCM+AC. When integrating the FCM clustering with an AC model, it is capable of targeting effective segmentation in medical imaging with messy or sparse annotations, hence its selection. The FCM+AC model's parameterization is crucial for its application to unsupervised segmentation tasks. Fuzziness is set at a level of 3 to allow for overlapping clusters, reflecting the inherent ambiguity in MRI textures. The algorithm segments images into three clusters, optimized for typical brain tissue contrasts seen in MRI scans. The active contour parameters include a window size of 7x7 to adjust contours closely around detected features, iterating 400 times to ensure precise adaptation to local variations. The length penalty of 0.000001 helps maintain the smoothness of the contour lines against pixel-level noise. An epsilon of 0.3 marks the threshold for convergence, balancing between algorithmic precision and computational demand.

4.3.3 Defining SVM. The SVM was implemented due to its robustness in high-dimensional data environments, such as our task in MRI imaging, where it can model the complex, non-linear relationships of our data efficiently. The SVM model is implemented with a Radial Basis Function (RBF) kernel to effectively handle the non-linearity in MRI data segmentation. The kernel's flexibility allows it to form a hypersurface in the high-dimensional space that can efficiently separate tumor from non-tumor regions based on feature similarities. The SVM parameters include C, the regularization parameter, which controls the trade-off between achieving a lower error on the training data and minimizing the model complexity for better generalization. The gamma parameter in the RBF kernel, which defines the influence of individual training examples, is tuned through a grid search approach to find the optimal balance between bias and variance. Feature scaling is applied to standardize the input feature vectors, essential for the kernel's performance, as it measures distances between vectors.

4.4 Evaluation Plan

4.4.1 Justification of Cross-validation Approach and Evaluation Metrics **Cross-validation Approach:** Ultimately the SVM and UNET are to be compared using 5-fold cross validation (CV), however the UNET is validated using a Single validation split strategy as of now (80:20) to reduce computation time and resources, whereas 5-fold CV is actively implemented in the SVM. Since FCM+AC is unsupervised, K-fold CV cannot be utilized as there is no ground truth for comparison. However, the dataset can still be manually split into subsets to simulate a consistency check without ground truths. Incorporating CV in UNET and SVM is encourages model robustness by utilizing different subsets of the dataset for training and validation. Consequently, mitigating overfitting and preventing any model dependency on a particular data subset.

Evaluation Metrics: The most popular and widely used performance metrics for brain tumor segmentation tasks include Sensitivity (Recall), Specificity, Accuracy, Precision, the Confusion matrix, Jaccard Index (IoU), and Dice Similarity [13], which we will likewise implement in our models as shown in Table **??**

The Dice and Jaccard indices are most appropriate for segmentation tasks, and UNET [16], as these metrics measure the overlap in pixels of the predicted masks and the ground truth annotations. Although similar to Dice, SVM will use the F1 score as it is primarily a classification metric and SVM is primarily a classifier [**?**]. However, since we have engineered the SVM to segment by classifying pixels, it will also share the Dice and IoU metrics, as well as Sensitivity, Specificity, Accuracy, and Precision. FCM+AC clustering quality of training will be evaluated via Silhouette Score, which indicates correct assignment, Davies-Bouldin Index, indicating similarity between clusters, and Calinski-Harabasz index, measuring the ratio of inter- to intra-cluster dispersion [11]. For testing evaluation, FCM+AR will be evaluated using the same metrics as UNET and SVM, because in conclusion each of the three models must be compared back to the ground truth

Table 1. Model Evaluation Metrics

SVM	UNET	FCM+AC
Sensitivity	Sensitivity	Sensitivity
Specificity	Specificity	Specificity
Accuracy	Accuracy	Accuracy
Precision	Precision	Precision
Confusion Matrix	Confusion Matrix	Confusion Matrix
Jaccard Index (IoU)	Jaccard Index (IoU)	
Dice Similarity	Dice Similarity	
F1 Score		

labels for evaluation, hence we can easily compare the models with these overlapping metrics. The shared success across the models will be evaluated in two stages. First, visual comparison will serve as a qualitative measure for the final predicted masks compared to the ground truth for each model. Second, we will compare the models according to these shared metrics.

4.4.2 Benchmarking Since the purpose of our benchmark dataset is to evaluate our model on unseen data, we do not plan to address data imbalances. Instead, we will preprocess the benchmark dataset using the same methods applied to our main dataset and then use our final models to generate predictions on our benchmark dataset and compute the same set of evaluation metrics utilized in our main dataset. Additionally, we will visualize our model predictions with ground truth annotations to qualitatively assess performance and pinpoint opportunities to improve our model. For a solid comparison, we will leverage visualization techniques such as ROC curves and heatmaps to compare the model's behavior across the seen (main dataset) and unseen (benchmark dataset) data. This comparison will enable us to evaluate our model and assess their generalization capacity to be refined as needed.

4.4.3 Data Augmentation Beyond integrating benchmarking to measure model robustness, we also made a custom data augmentation algorithm [see GIT], and compared how well a model could perform on unseen data had it been either trained with data augmentation or not. The image data augmentation generator built from Keras, has variable parameters to allow for realistic, but potentially challenging image variations for the model to learn from and build robustness.

5 Results

We evaluated our model's performance in the following ways: (1) through the use of visualizations for qualitatative checks and (2) by generating evaluation metrics. In this section we present the results of these evaluations.

5.1 Individual Results

5.1.1 SVM During training we evaluated the performance of different kernels during cross validation of our dataset. Additionally, we used grid search to determine the best hyper-paramaters to include in our model. Unfortunately, this proved too computationally expensive given our resources, so we moved forward with a predetermined regularization parameter, C = 1.0. Figure 4 demonstrates our results. From our results we took our best performing kernel, RBF and used it to finalize our model. This model was then used to test our final dataset. The performance of our final model on the testing data is visualized in Figure 5.

To assess performance per tumor, We generated a confusion matrix to demonstrate model performance across different tumor types 9. Additionally, we visualized our overall model performance as seen in Figure 6 in Appendix. Unfortunately we were unable to generate results for our augmented dataset due to difficulty implementing augmentation in the SVM pipeline.

Table 2. Evaluation Metric Comparison across Model

Metric	SVM	UNET	Sigmoid	Average
Accuracy	0.92	Data	Data	Data
Precision	0.92	Data	Data	Data
Recall	0.92	Data	Data	Data
Jaccard Index	0.85	Data	Data	Data
F1 Score	0.92	Data	Data	Data
Dice Similarity	0.97	Data	Data	Data
Specificity	Data	Data	Data	Data

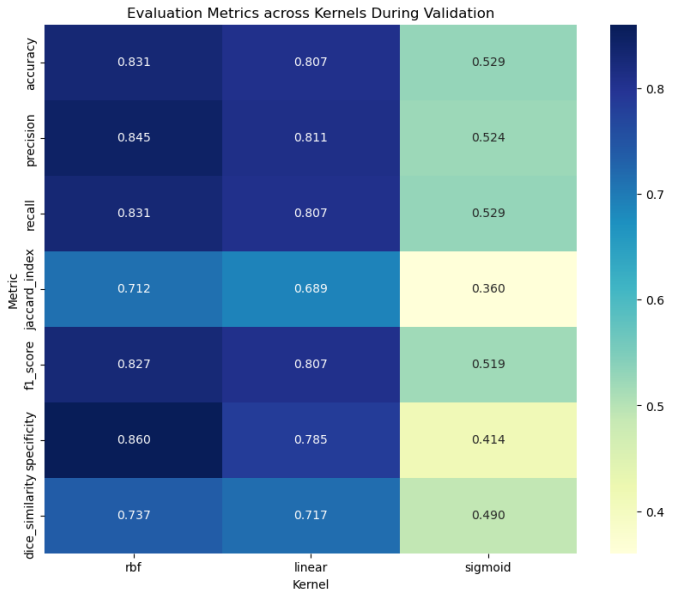

Fig. 4. Heatmap showing evaluation metrics by kernel type generated during training [4]

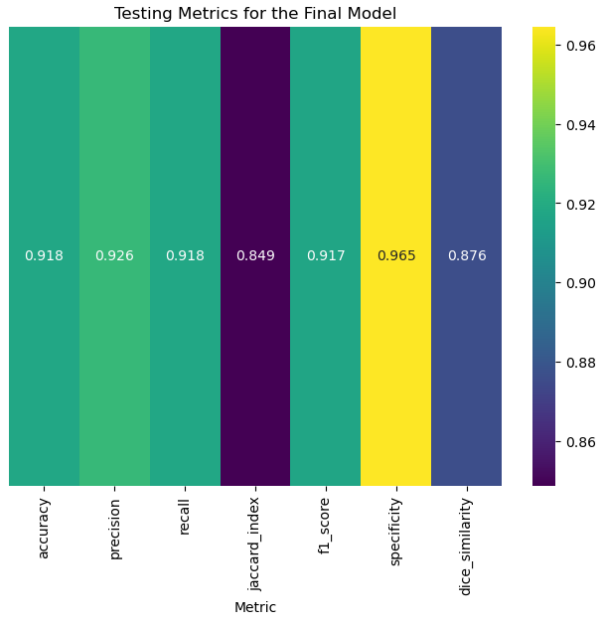

Fig. 5. Evaluation Metrics when apply RBF kernel and C=1.0 to testing data. [4]

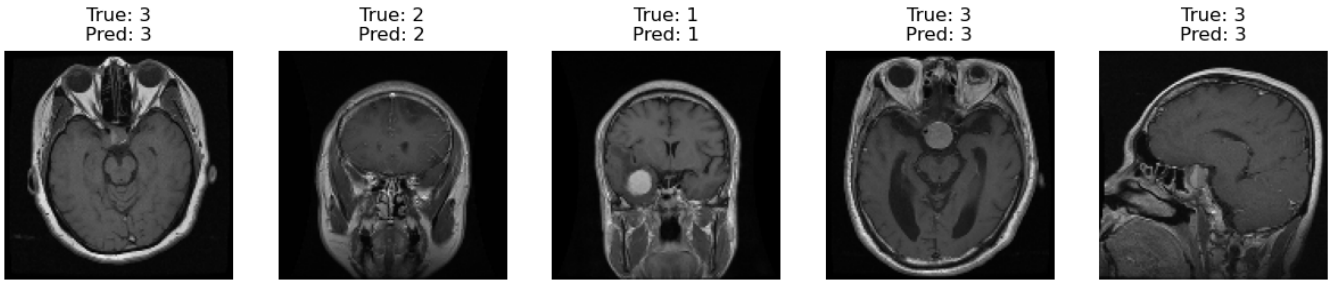

Fig. 6. Visualization of Result on Testing Set. [4]

6 Discussion

SVM performance. The performance evaluation of the SVM model using the RBF kernel and regularization parameter of C = 1.0 provides insight into its effectiveness in MRI image segmentation. During training the model achieved an average accuracy of 83.05% indicating its ability to correctly predict class labels for the training dataset. The precision and recall metrics 84% and 83% respectively show that the model was able to identify postivie instances. Whereby the Jaccard Index, measuring similarity between predicted and actual labels generated a moderate score of 71.22%. The F1 and Dice similarity score indicate the model was fairly balanced in its performance. N

During testing we see that model's accuracy increased, with a performance of 91.76% indicating a strong potential for generalization. Precisoin and recall were also notably high, average 92%.The Jaccard and Dice scores also demonstrate strong correlation between true and predicted values. The remaining metrics also boast scores of 90% or greater with the exception of the Jaccard index scoring 86% indicating that the model was slightly less effective in identifying negative instances.

UNETT performance. The UNET results indicate strengths and opportunities for improvement. During training, the model achieved a high accuracy of 99% significantly outperforming the SVM. The recall and specificity score were also very high indicating the model's excellence in in capturing positive and negative instances. The IoU and Dice coefficient scores were considerably low (0.016 and 0.031 respectively), indicating poor performance in judging overlap and similarity between predicted and ground truth masks.This discrepancy between high accuracy and low IoU/ Dice scores suggesting potential challenges in accurately identifying object boundaries, hinting at potential challenges iat performing segmentation tasks.

6.1 Study Limitations

Throughout this study we encountered several limitations which we describe below:

- UNET overfitted, with limited time we were unable to address this and test out different hyperparamaters
- Limited computational power meant we were unable to use 512 x 512 dimensions allow for more accurate tumor identification
- Limited computational power inhibited data augmentation efforts
- Challenges with getting our dataset to match the Cheng dataset resulted in poor benchmarking results

6.2 Future Work and Applications

Throughout this work, we became aware of a number of opportunities to continue in future directions with segmentation and deepen the exploration of these ML/DL methods in later works. Potential future directions include:

Venue a) Metrics Per Tumor Type: To deepen our understanding of model performance, we could explore detailed metrics for each tumor type. This will allow us to tailor our models more precisely to the unique challenges posed by the three different tumor morphologies and their respective properties. Venue b) Using High-Quality Images: Recognizing the potential the models, especially the UNET has could refine our model architecture's to specifically leverage images with a resolution of minimum 512x512 pixels or higher. This could ensure that the models can capitalize on the detailed information available in these higher-quality scans and learn more effectively, as lower resolution indicated worse training. Venue c) Overcoming Computational Limitations: To address the substantial computational demands of advanced deep learning, our immediate next step would be to connect to a cluster equipped with NVIDIA GPUs, such as the BMI department cluster. While having experimented with this cluster, we have yet to run the full pipeline on it. This will allow for potentially more complex model training and 5-fold validation in UNET. However, our experimentation indicated memory on the cluster was an issue when implementing .ipynb, and therefore operating at the command line and alleviating image expansion will be key. Venue d) Flagging Unlabeled or Unrecognized Tumors: While our model was trained and benchmarked on datasets that were encoded with labels for only 3 tumors, it would be beneficial to include an algorithmic component to identify unrecognizable tumors in the case of bench marking on data with other tumor types. Additionally, this would help to ensure overfitting is not occuring, as the model would fail its testing if it's only memorized three tumor types. Venue e) Generalizability on CE vs. Non-CE Images: Our current focus on non-contrast enhanced (non-CE) MRI images was to assess model performance consistently across a single modality. Expanding the benchmark to include contrast-enhanced (CE) images will test the model's robustness and real-world applicability, ensuring it can handle diverse and complex datasets typically found in medical facilities.

7 Conclusion

Our comparative study demonstrates the distinct capabilities and limitations of SVM, UNET, and FCM+AC models in the context of brain tumor MRI image segmentation, to enhance the interpretative facets of medical image analysis. Through rigorous evaluation, including cross-validation and various performance metrics, we reinforce the necessity for adaptive model selection based on the specific data characteristics of MRI and computational constraints. Moving forward, the insights garnered here will inform the optimization and application of these methodologies, aiming to improve both the accuracy and efficiency of tumor diagnosis in clinical settings.

8 Appendices

8.1 Figure for Background and Model Selection

8.2 SVM Evaluation Results

8.3 Metric Calculations (Detailed Vers.)

Metrics used for UNET and SVM

Explanation of terms:

- TP: true positive
- FP: false positive
- TN: true negative
- FN: false negative
- IoU: Intersection over Union

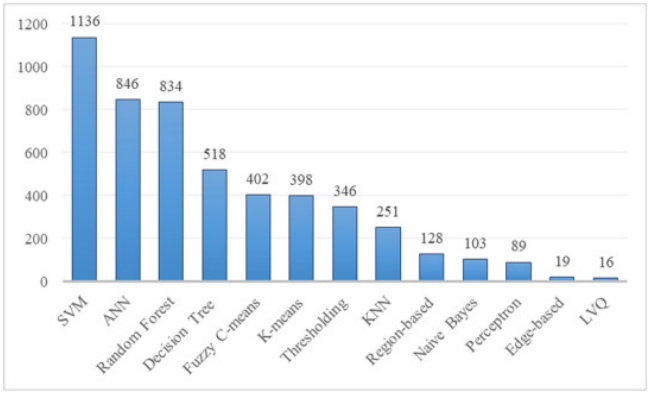

Fig. 7. Number of publications between 2015 and 2022 for brain tumor segmentation using supervised and unsupervised learning techniques. [13]

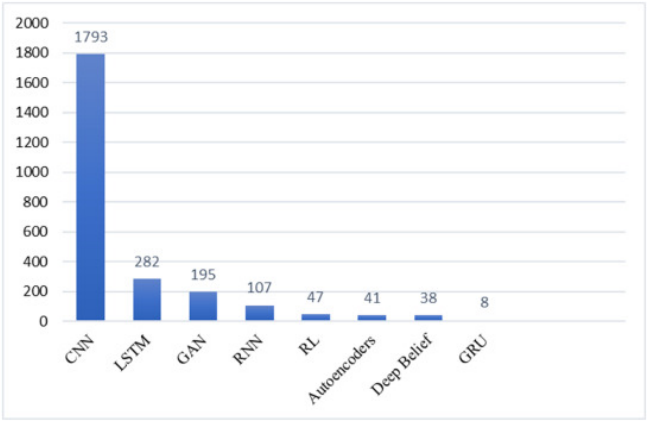

Fig. 8. Number of publications between 2015 and 2022 for brain tumor segmentation using DL models. [13]

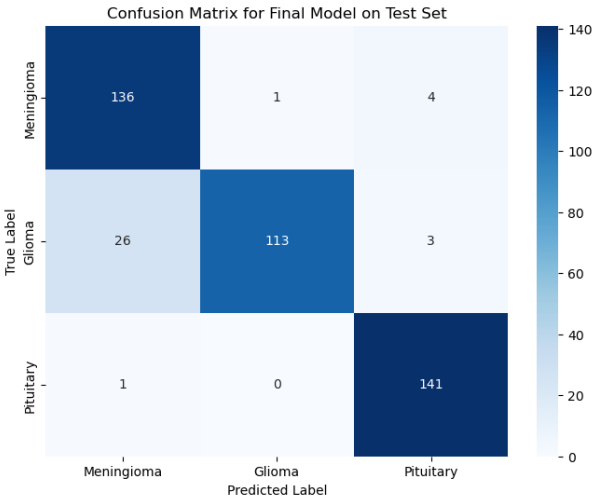

Fig. 9. Confusion matrix showing overall model performance by tumor type. [4]

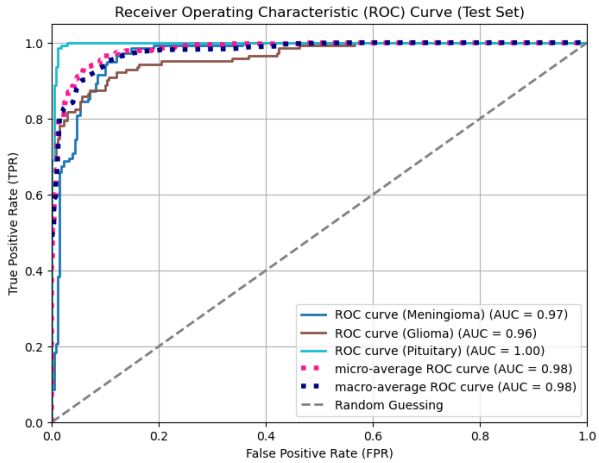

Fig. 10. ROC Curve for testing data. [4]

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}},$$

This metric allowed us to evaluate to proportion of true positive cases (correctly identified tumor regions). High sensitivity meant fewer false negatives.

$$\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}},$$

Specificity allowed us to evaluate the proportion of true negative cases (healthy regions) correctly identified by the model. Contrasting sensitivity, a high specificity indicated fewer false positives.

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}$$

This metric allowed us to measure the overall correctness of the model across all classes by comparing the true labels to the predicted labels.

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

This metric measured the accuracy of positive predictions made by the model. In other words, it evaluates the degree to which the model is capable of not labeling a negative sample as positive. It is calculated by taking the ratio of true positives and false positives. The best value is 1 and the worst value is 0.

$$\text{Confusion Matrix} = \begin{bmatrix} \text{TN} & \text{FP} \\ \text{FN} & \text{TP} \end{bmatrix},$$

The confusion matrix evaluates the accuracy of the predictions per class. In other words, it presents the prediction summary in matrix form.

$$\text{Jaccard Index (IoU)} = \frac{\text{TP}}{\text{TP} + \text{FP} + \text{FN}}$$

Also called the Jaccard similarity coefficient, this metric is used to compare a set of predicted labels to their true labels.

$$\text{Dice Similarity} = \frac{2 \times \text{TP}}{2 \times \text{TP} + \text{FP} + \text{FN}}$$

This metric is used to gauge the similarity between two sets of data.

Metrics used for Fuzzy C-Means (FCM)

The following metrics assess cluster quality and separation.

Explanation of terms:

- N: total number of data points in dataset
- *b(i)*: the average distance between data point i and all other data points in the nearest neighboring cluster
- *a(i)* the average distance between a data point i and all other data points within its own cluster
- k: Number of clusters
- σ_i: Average distance from a datapoint i to all other data points in the same cluster
- $d(c_i, c_j)$: Distance between centroids c_i and c_j of clusters i and j

$$\text{Silhouette Score} = \frac{1}{N} \sum_{i=1}^{N} \frac{b(i) - a(i)}{\max(a(i), b(i))},$$

This score measures how well-separated clusters are.The overall silhouette score is computed as the average of all data points. Higher scores indicate well-clustered data points.

$$\text{Davies-Bouldin} = \frac{1}{k} \sum_{i=1}^{k} \max_{j \neq i} \left(\frac{\sigma_i + \sigma_j}{d(c_i, c_j)} \right)$$

This metric evaluates cluster separation and compactness. Lower values mean better clustering.

$$\text{Calinski-Harabasz} = \frac{\text{Tr}(B)}{\text{Tr}(W)} \times \frac{N - k}{k - 1}$$

With this metric we assess the ratio of between cluster dispersion to within-cluster dispersion. Higher values indicate better-defined clusters.

8.4 Individual Results

8.4.1 SVM

8.4.2 FCM

8.4.3 UNET

8.4.4 Comparison across models

8.4.5 Comparison with literature

References

[1] Magnetic resonance imaging (mri).

[2] Jun Cheng. Brain tumor dataset. figshare. Dataset, 1512427(5), 2017.
[3] Jun Cheng, Wei Huang, Shuangliang Cao, Ru Yang, Wei Yang, Zhaoqiang Yun, Zhijian Wang, and Qianjin Feng. Enhanced performance of brain tumor classification via tumor region augmentation and partition. PLOS ONE, 10(10):e0140381, October 2015.
[4] Guang Feng and Chong Tang. Portrait semantic segmentation method based on dual modal information complementarity. Applied Sciences, 14(4):1439, 2024.
[5] Chris Giordano, Meghan Brennan, Basma Mohamed, Parisa Rashidi, François Modave, and Patrick Tighe. Accessing artificial intelligence for clinical decision-making. Frontiers in digital health, 3:645232, 2021.
[6] Haichun Li, Ao Li, and Minghui Wang. A novel end-to-end brain tumor segmentation method using improved fully convolutional networks. Computers in Biology and Medicine, 108:150–160, May 2019.
[7] Xuhong Li, Haoyi Xiong, Xingjian Li, Xuanyu Wu, Xiao Zhang, Ji Liu, Jiang Bian, and Dejing Dou. Interpretable deep learning: Interpretation, interpretability, trustworthiness, and beyond. Knowledge and Information Systems, 64(12):3197–3234, 2022.
[8] Benjamin Maas, Erfan Zabeh, and Soroush Arabshahi. Quicktumornet: Fast automatic multi-class segmentation of brain tumors. In 2021 10th International IEEE/EMBS Conference on Neural Engineering (NER), pages 81–85, 2021.
[9] Mamta Mittal, Lalit Mohan Goyal, Sumit Kaur, Iqbaldeep Kaur, Amit Verma, and D. Jude Hemanth. Deep learning based enhanced tumor segmentation approach for mr brain images. Applied Soft Computing, 78:346–354, May 2019.
[10] Geethu Mohan and M. Monica Subashini. Mri based medical image analysis: Survey on brain tumor grade classification. Biomedical Signal Processing and Control, 39:139–161, January 2018.
[11] Samina Naz, Hammad Majeed, and Humayun Irshad. Image segmentation using fuzzy clustering: A survey. In 2010 6th international conference on emerging technologies (ICET), pages 181–186. IEEE, 2010.
[12] Wynton B. Overcast, Korbin M. Davis, Chang Y. Ho, Gary D. Hutchins, Mark A. Green, Brian D. Graner, and Michael C. Veronesi. Advanced imaging techniques for neuro-oncologic tumor diagnosis, with an emphasis on pet-mri imaging of malignant brain tumors. Current Oncology Reports, 23(3):34, February 2021.
[13] Ramin Ranjbarzadeh, Annalina Caputo, Erfan Babaee Tirkolaee, Saeid Jafarzadeh Ghoushchi, and Malika Bendechache. Brain tumor segmentation of mri images: A comprehensive review on the application of artificial intelligence tools. Computers in Biology and Medicine, 152:106405, January 2023.
[14] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18, pages 234–241. Springer, 2015.
[15] S. Srakocic. Can an mri detect cancer? what it can and can't detect, Mar 2023.
[16] S Kevin Zhou, Hayit Greenspan, Christos Davatzikos, James S Duncan, Bram Van Ginneken, Anant Madabhushi, Jerry L Prince, Daniel Rueckert, and Ronald M Summers. A review of deep learning in medical imaging: Imaging traits, technology trends, case studies with progress highlights, and future promises. Proceedings of the IEEE, 109(5):820–838, 2021.