3D Brain Tumor Segmentation and Visualisation for Improved Clinical Understanding

Ershveer Singh Saini
240131632
MSc.Artificial Intelligence
e.saini@se24.qmul.ac.uk
Kit Mills Bransby
Queen Mary University of London

***Abstract*—Deep learning models can achieve high accuracy in brain tumor segmentation, but their decision-making process often remains opaque to clinicians. This project addresses the challenge of making segmentation outputs more interpretable and clinically useful through advanced visualization techniques. A 3D U-Net was trained on the BraTS2021 multimodal MRI dataset to segment gliomas and their subregions, forming the foundation for a visual explanation. Incorporate Grad-CAM heatmaps to highlight spatial regions influencing predictions, voxel-level Softmax probability maps to show class confidence, and Monte Carlo Dropout-based variance maps to quantify uncertainty, and an interactive 3D web application to visualise the tumor section using the Marching Cubes algorithm, allowing clinicians to explore tumor morphology, assess model confidence, and understand prediction rationale. This work demonstrates that visualising predictions, uncertainty, and network attention can bridge the gap between deep learning performance and clinical interpretability, supporting more informed decision-making in neuro-oncology.**

I. Introduction

Deep learning segmentation models for brain tumors have achieved remarkable accuracy, yet their 'black-box' nature poses challenge for clinical use (Teng et al., 2022; Salahuddin et al., 2021) Clinicians require not only accurate predictions, but also understandable explanations of how and why a model reached its conclusions, especially in high stakes settings like neuro-oncology. Visualisation tools such as heatmaps and uncertainty overlays serve as interpretable bridges between complex model behaviors and human intuition. By making predictions, model attention, and confidence transparent, these visualisations foster clinicians' trust, support diagnostic reasoning, and facilitate actionable insights (Hanif et al., 2021; Chen et al., 2022). Therefore, the focus of this work is not solely segmentation, but rather how to interpret and communicate model outputs effectively to clinicians through visualisation. Visual representations such as saliency heatmaps and uncertainty overlays make AI decisions more interpretable and actionable for clinicians. Attribution-based tools like Grad-CAM, integrated gradients, and maps highlight image regions influencing model outputs, enhancing localisation and interpretability (Alam et al.,2025; Huff et al., 2019). Meanwhile, visualising uncertainty through methods like Monte Carlo Dropout helps clinicians understand the model's confidence, prioritize review of ambiguous areas, and avoid overconfidence in automated outputs. (Lambert et al.,2022).

In light of these motivations, this project leverages segmentation not as an end goal but as a foundation for developing interpretable visualisation tools, as shown in Figure 1. The core is to translate AI outputs into forms clinicians can visually inspect through Grad-CAM heatmaps, voxel-wise softmax probability, uncertainty visualisation, and a 3D web application.

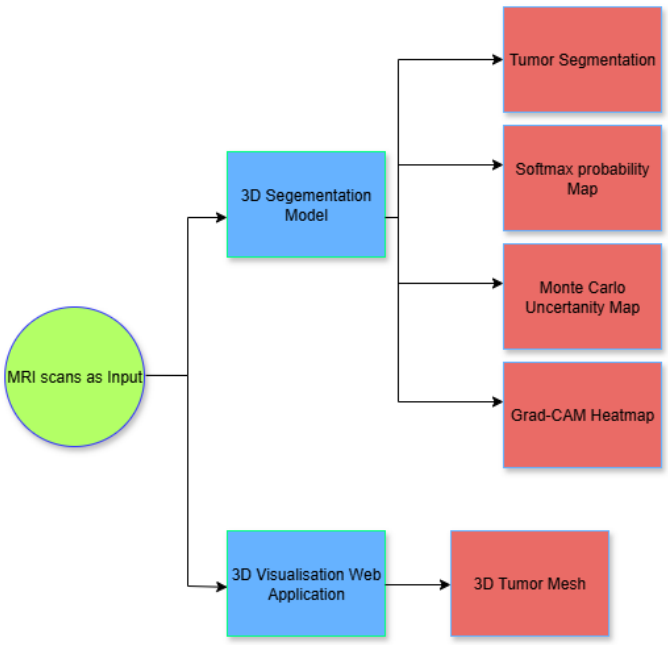

Fig. 1. Project Overview

II. Related Work

The segmentation of brain tumors from medical images has been a significant area of research for decades. While early methods relied on traditional image processing techniques such as thresholding and region growing, the field was revolutionsed by the advent of deep learning, particularly Convolutional Neural Networks (CNNs). A seminal breakthrough in biomedical image segmentation was the introduction of the

U-Net (Ronneberger et al., 2015). This architecture introduced a now-iconic U-shaped structure, consisting of a contracting path (encoder) to capture context and a symmetric expanding path (decoder) to enable precise localisation. The key innovation was of skip connections, which concatenate high-resolution feature maps from the encoder with upsampled outputs of the decoder. This fusion of high-level contextual information with fine-grained spatial detail proved remarkably effective for segmentation tasks, allowing the network to achieve high accuracy even with very few training samples. While the original U-Net designed for 2D images, the volumetric nature of MRI data necessitated a transition to three-dimensional models (Çiçek et al., 2016), directly extended the original architecture by replacing all 2D operations with their 3D counterparts, creating the 3D U-Net. This allowed the model to process entire volumes at once, capturing the crucial inter-slice connectivity and full spatial context of tumors, a significant advantage over by slice-by-slice 2D approach. Concurrently, (Milletari et al., 2016) proposed the V-Net, another 3D fully convolutional network that shared the U-net's encoder-decoder structure. A key contribution of the V-Net was the introduction of a novel loss function based on the Dice coefficient, which is particularly well suited for handling the severe class imbalance often present between the small tumor regions and the large background in medical scans. The culmination of these advancement is arguably represented by the nnU-Net framework, (Isensee et al., 2021), rather than proposing a single new architecture, nnU-Net is a self configuring framework that automatically adapts the U-Net architecture, preprocessing steps, and training parameters to the specific properties of any given dataset. By systematically applying a set of heuristic rules and empirical choices, nnU-Net consistently achieves state of the art performance across a wide variety of segmentation tasks, including the BraTS challenge, establishing a powerfull and robust baseline for the field. Beyond achieving high segmentation accuracy, a critical aspect of translating deep learning models into a clinical setting is establishing trust and understanding their decision-making process. One of the most influential techniques for visualising model attention is Gradient-weighted Class Activation (Grad-CAM), introduced by Selvaraju et al. (2017). Grad-CAM produces a coarse localization map that highlights the specific regions in the input image that were most important for a particular prediction. In parallel, quantifying a model's confidence in its prediction is crucial for high-stakes applications like medical diagnosis. A groundbreaking approach for estimating this uncertainty is Monte Carlo (MC) dropout, proposed by Gal and Ghahramani (2016). They demonstrated that by keeping dropout active during inference and performing multiple forward passes, dropout can be used as a practical approximation of Bayesian inference in deep neural networks. In addition, research has also focused on creating tools for visualising 3D medical data. For example, (Hähn et al., 2012) developed Slice: Drop, a web application for viewing 3D MRI volumes with interactive features like opacity and thresholding. While powerful, such general-purpose tools are often not directly integrated with a specific model's analysis pipeline. This highlights the opportunity to create a custom visualisation tool tailored to the outputs of the segmentation model developed in this project. Building upon the success of the 3D U-Net architecture, this dissertation implements a custom model tailored to the BraTS 2021 dataset. Furthermore, it extends the standard segmentation task by incorporating an analysis of model interpretability and uncertainty using Grad-CAM, SoftMax probability heatmaps, and Monte Carlo Dropout, as well as a 3D visualization app, which are critical for building clinical trust. Additionally, it develops a novel 3D interactive visualization tool.

III. Methodology

This section outlines the dataset used, the data preprocessing techniques applied, and the architectures of the 3D U-Net implemented for the image segmentation.

A. BraTS Dataset

The Brain Tumor Segmentation (BraTS) challenge, held annually at the MICCAI (Medical Image Computing and Computer Assisted Intervention) conference since 2012, provides a benchmarking platform for developing and evaluating automated brain tumor segmentation methods using state-of-the-art deep learning (Menze et al., 2015; Bakas et al., 2018).

The BraTS 2021 dataset consists of multimodal 3D MRI scans from glioma patients, collected across multiple institutions under varying imaging protocols, ensuring realistic heterogeneity (Bakas et al., 2017). Each case includes four MRI sequences- T1, T1-contrast (T1ce), T2, and FLAIR, along with the expert-annotated ground truth labels for tumor subregions (necrosis, edema, and enhancing tumor). Data are provided in NIfTI format as volumetric grids of voxels, a voxel is the 3D equivalent of a pixel, representing a small cube of tissue in volumetric imaging, unlike a 2D photograph, which has height and width, this data also has depth, allowing for a full, three-dimensional representation of the brain and tumor(Datature, 2025); making the dataset a standard benchmark for developing and comparing brain tumor segmentation algorithms.

For each patient, five distinct 3D volumetric MRI scans are provided. T1-weighted (T1): Provides detailed anatomical information. T1-weighted with contrast enhancement: Uses a contrast agent to highlight the active, enhancing parts of the tumor. T2-weighted (T2): Is particularly sensitive to areas with high water content, clearly showing the edema (swelling) around the tumor. Fluid Attenuated Inversion Recovery (FLAIR): Like T2 but suppresses the signal from cerebrospinal fluid, making the edema even more conspicuous and the Segmentation file.

B. Data Preprocessing

The MRI scans and their corresponding segmentation masks, which are in the NIfTI format, were loaded using the nibabel library. The intensity values of the MRI images

(Flair, T1ce, T2) were scaled using MinMaxScaler from scikit-learn. Each modality was reshaped to a 2D array, scaled between 0 and 1. This step helps in standardizing the intensity range across different scans. The segmentation masks have classes represented by values 0,1,2, and 4. To work with a consistent range of class indices for one-hot encoding and model output (0 to 3), the class value 4 was remapped to 3. The images and masks were centrally cropped to a uniform size of (128,128,128). This step is important for handling variations in image dimensions in the dataset and reducing computation load. The crop was applied to combined image modalities and to the segmentation mask. Volumes with very small tumor regions were filtered out. The was performed by checking the unique values and their counts in the cropped segmentation mask. If the proportion of non-background pixels was below a certain threshold (0.01), the sample was ignored during the saving process.

The three pre-processed modalities for each patient were stacked along a new dimension to create a multi-channel input tensor. The shape of this stacked tensor became (D,H,W,C), depth, height, width, channels, eg.(128,128,128,3). The pre-processed and cropped segmentation masks were converted from indexed class labels (0,1,2,3) into a one-hot encoded format. This results in a mask tensor of shape (D,H,W,C), depth, height, width, number of classes, eg.(128,128,128,4). The preprocessed image volumes and their corresponding one-hot encoded masks were saved.The saved preprocessed data was split into training, validation, and test sets with a ratio of 75% for training, 15% validation, and 10% test. BratsDataset and DataLoader were implemented to efficiently load the preprocessed numpy files during training, validation, and testing. This custom dataset handles loading the images and mask files, converting them to PyTorch tensors, permuting the axes to the required format for 3D CNNs, and applying

C. 3-D Segmentation

The 3D U-net is a convolutional neural network designed specifically for volumetric medical image segmentation, extending the original 2D U-Net by incorporating a third spatial dimension to capture contextual information across MRI slices (Çiçek et al., 2016). Its structure follows, as shown in Figure 2, a characteristic encoder-decoder design with skip connections:

Encoder: progressively downsamples the 3D volume, reducing spatial resolution while increasing the number of feature maps to capture an increasingly abstract representation of tumor structure.

Bottleneck Layer: the central layer where the network encodes the most compressed representation of the input volume, retaining rich semantic features.

Decoder: upsamples the feature amps back to the original spatial resolution. Skip connections link encoder layers to decoder layers, ensuring that fine-grained spatial details are preserved while reconstructing the segmentation mask. The output layer produced a multi-channel 3D segmentation mask, where each channel corresponds to a predicted tumor subregion.

The forward method features how the input tensor flows through the network. It shows the sequence of operations, including the down-sampling in the encoder, the bottleneck, the up-sampling and the skip connection in the decoder, and the final output convolution.

Then, initiate the U-Net 3D model with the 3 input channels (for the MRI modalities) and 4 output channels (for the classes). It then creates a dummy input tensor with a batch size of 1 and the expected dimensions. The input tensor is passed through the model to perform a forward pass, and the shape of the input and output tensors is printed. This is a simple check to ensure the model is structured correctly and the tensor dimensions are as expected. The output shape confirms that for a single input volume with 3 channels and dimensions 128x128x128, the model outputs a volume of the same spatial dimensions but with 4 channels, corresponding to the predicted logits for each class at every voxel.

IV. VISUALISATION

This section describes the visual analysis used to evaluate model performance, highlighting segmentation quality, model confidence, and regions of focus during prediction. Four types of visualisations were generated for each patient: 3D mesh model using Marching Cubes, voxel-wise Softmax probability maps. Monte Carlo Dropout uncertainty maps, and Grad-CAM heatmaps. The figures present five representative slices per anatomical (axial, coronal, sagittal), with colour legends and bars included to aid interpretation of tumor regions, probability values, and uncertainty levels.

- 3D Visualization Web Application: The primary purpose of this application is to provide a dynamic and intuitive tool for the qualitative analysis of the BraTS2021 dataset. Instead of viewing 2D slices, this tool reconstructs the tumor and surrounding brain anatomy as a complete 3D model, allowing for a more holistic understanding of the tumor size, shape, and location. The application is built using the Flask web framework, which runs on a local web server to handle data processing and render the final visualization in a standard web browser.
  The application operates using a multi-page structure, with two main components or routes: a homepage, in Figure 3, for patient selection and another dedicated page for the 3D visualization, in Figure 5. The homepage when the user first navigates to the web server's address, this function is executed. Its job is to automatically scan the specified data directory and identify all the patient subfolders. It then generates a simple, user-friendly homepage containing a dropdown menu with all the patient IDs. This page allows the user to easily select any patient from the dataset for visualization. Next, on the visualization page, when the user selects a patient and clicks 'Visualize', the browser is directed to a unique URL for that patient. This triggers the main visualization pipeline. The script loads the necessary NIfTI files for the selected patient, the ground truth segmentation, and the T1-weighted anatomical scan. The core of the 3D

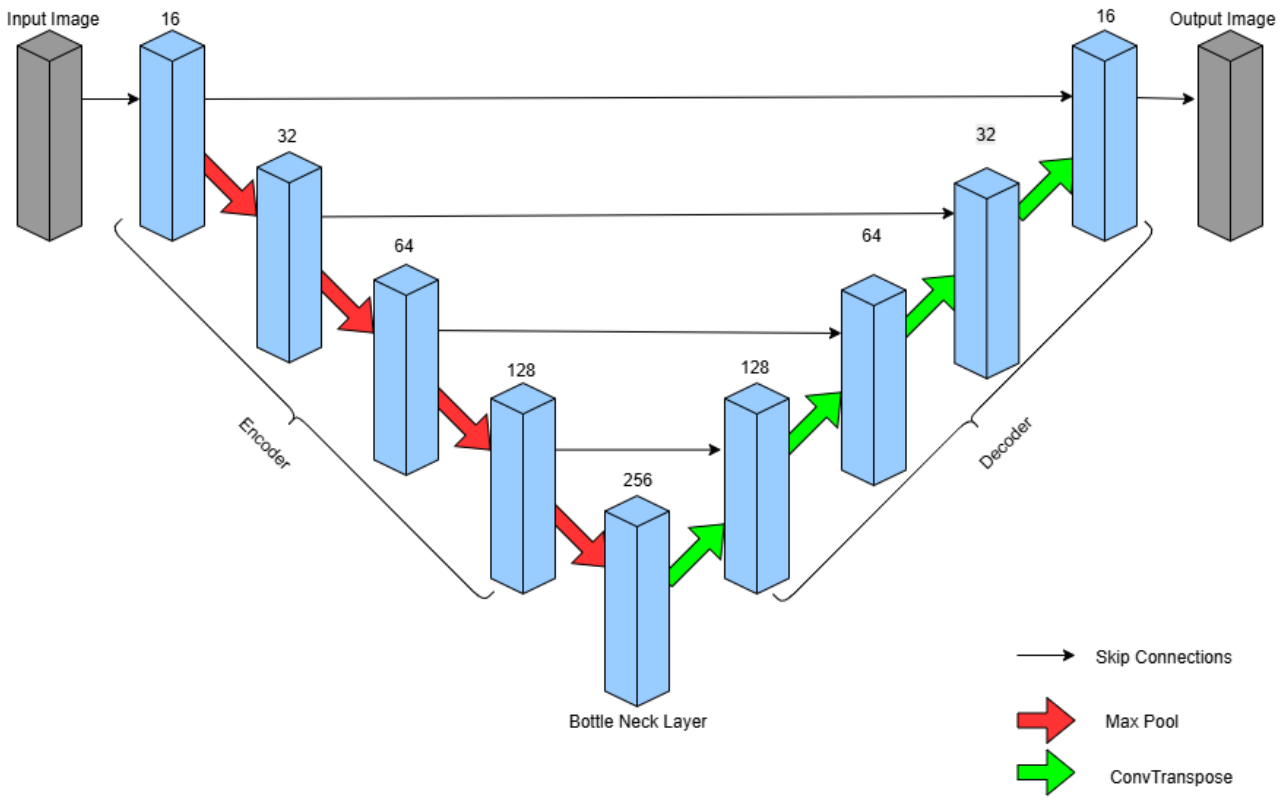

Fig. 2. 3D U-Net Architecture

reconstruction happens here, as visible (zoomed tumor image) in Figure 4. The script processes both the T1 scan and the segmentation mask to create 3D surface models using the Marching Cube algorithm, which is a computer graphics method for extracting a polygonal mesh of an isosurface from three-dimensional volumetric data (Lorensen & Cline, 1987). It generates a separate mesh for the brain surface and for each tumor (Necrotic Core, Edema, and Enhancing Tumor) based on their labels in the segmentation mask. Using the Plotly library, an open-source graphic and data visualisation library (Inc., 2015), the script constructs a 3D scene, and each generated mesh is added to the scene as an object, with specific properties for colour, opacity, and lighting to ensure a clear and informative visual. Finally, the complete interactive figure is converted into a self-contained block of HTML and JavaScript. This HTML is then sent to the user's web browser, which renders the final, interactive 3D plot.

The visualization, in the Figure 4, shows a semi-transparent, gray 3D model of the patient's brain, generated from the T1 MRI scan, the different tumor components are rendered as distinct, coloured 3D models, as you can see in the figure: the Necrotic Core in solid red, the Enhancing Tumor in semi-transparent blue, and the surrounding Edema in highly transparent green. The user has full control over the 3D scene; they can rotate, pan, and zoom to view the tumor from any angle. A set of

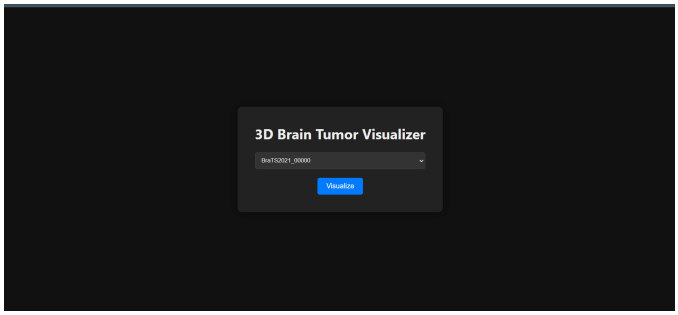

Fig. 3. Frontend of the web app

buttons at the top of the plot provides further control, allowing the user to toggle the visibility of specific components, such as viewing the 'Tumor Only' without the brain or isolating a single region like the 'Necrotic Core'.

- Softmax probabilities: represent the model's confidence for each class at each voxel. The Softmax function is applied to the raw neural network outputs at the end of the model and converts them into a probability distribution. For every voxel in the 3D scan, the Softmax ensures that each class's scores are between 0 and 1 (Number Analytics, 2025; Mackinlay, 2024) and that all class probabilities sum to 1. The equation shows how the softmax function converts raw logits (z_k) into a probability distribution

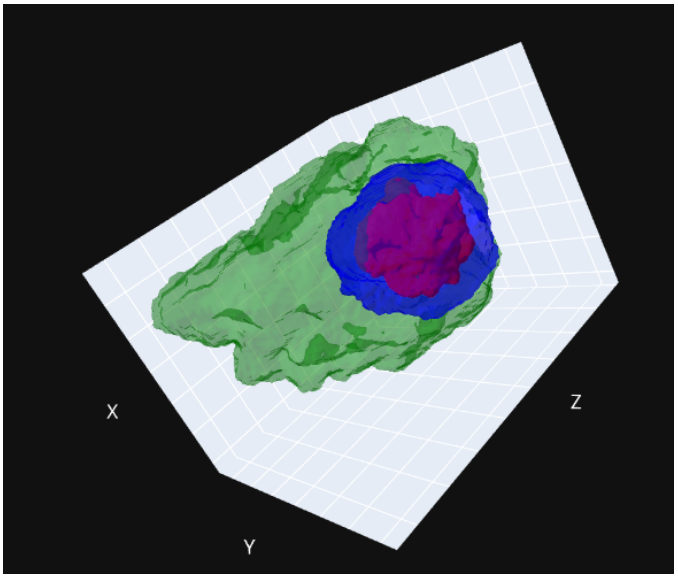

Fig. 4. 3D tumor mesh of patient ID: BraTS2021_00281

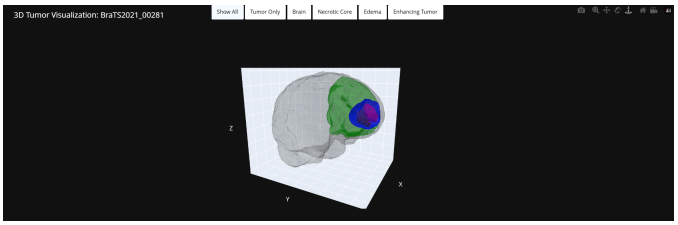

Fig. 5. Whole 3D brain-tumor mesh of patient ID: BraTS2021_00281, with different functionalites (buttons, on top)

across (K) classes, ensuring that each voxel's score sums to one.

$$P(y = k \mid \mathbf{z}) = \frac{e^{z_k}}{\sum_{j=1}^{K} e^{z_j}}$$

In Figure 6, in the Softmax probability map, on the predicted images, for each of the four classes (Background, Necrotic core, Edema, Enhancing Tumor, order-wise), for selected slices, allows us to see the model's confidence distribution across different regions and classes. A bright colour (like yellow or red in the 'hot' colourmap) indicates a high probability for that specific class at that voxel, meaning the model is more certain about that classification. Conversely, a dark colour (like black or dark red) indicates a low probability for that class at that voxel, meaning the model is less certain that the voxel belongs to that class. A higher intensity value (brighter colour) in a specific pixel of the probability map for a class indicates a higher likelihood that the corresponding voxel in the original image belongs to that class.

- Monte Carlo Dropout: is a widely used technique to estimate model uncertainty in deep learning by enabling dropout layers during inference and performing multiple stochastic forward passes (Gal & Ghahramani, 2016).

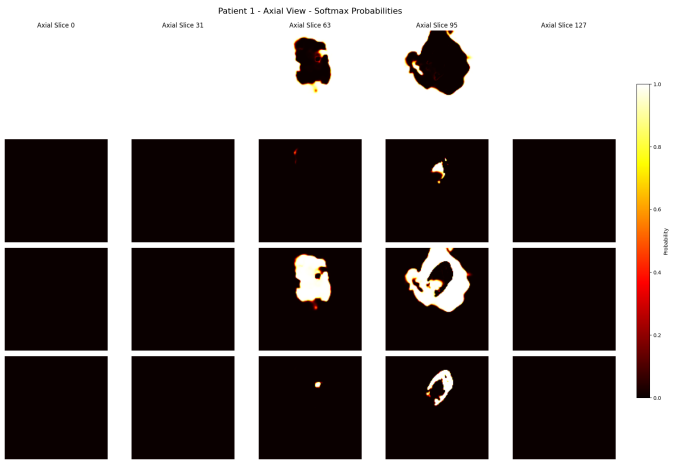

Fig. 6. Softmax Probability Map, for each class (Background, Necrotic core, Edema, and Enhancing Tumor), row-wise respectively, showing 5 slices [0,31,63,95,127], column-wise, for the axial plane, slice 95 shows the best tumor view, bright regions (yellow) represent high probability and strong model confidence, while darker regions (black) indicate lower probability and greater uncertainty about voxel belongs to that class.

Each pass effectively samples a thinned version of the network, producing slightly different predictions. Averaging across predictions yields mean probabilities, while the variance provides an estimate of uncertainty, which is particularly valuable in safety-critical domains such as medical imaging (Kendall & Gal, 2017; Abdar et al., 2021). Each forward pass with dropout enabled will effectively use a slightly different thinned version of the network due to randomly dropping out neurons. This results in slightly different predictions. The first equation estimates the mean probability by averaging predictions across N stochastic forward passes with dropout enabled, while the second equation uses the variance across passes to approximate the model's uncertainty. Where $p_k(x, z_n)$ is the predicted probability for class k in the n-th pass, and z_n is the dropout mask.

$$\bar{p}_k(\mathbf{x}) = \frac{1}{N} \sum_{n=1}^{N} p_k(\mathbf{x}, \mathbf{z}_n)$$

$$\text{Var}_k(\mathbf{x}) \approx \frac{1}{N} \sum_{n=1}^{N} \left(p_k(\mathbf{x}, \mathbf{z}_n) - \bar{p}_k(\mathbf{x})\right)^2$$

As in Figure 7, bright color (yellow) indicates high variance. Dark color (red) indicates low variance. So, Low variance is generally a good sign because it means the model's predictions for that region are consistent across different dropout masks, which indicates higher confidence in the prediction for the target class (Necrotic core, Edema, Enhancing Tumor), visualizing uncertainty. High variance (bright colors) shows that the model's predictions for that region fluctuate significantly across runs, suggesting lower confidence or greater uncertainty about the classification of those pixels for the target

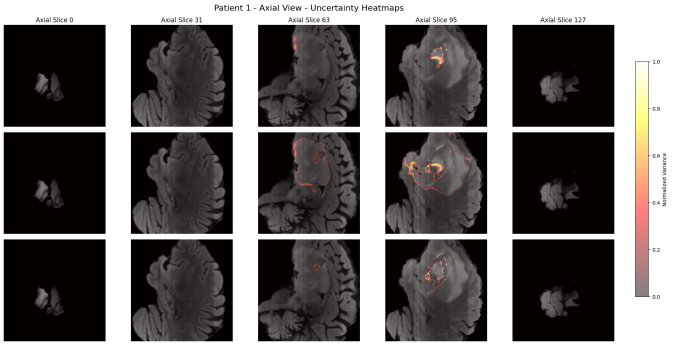

Fig. 7. Monte Carlo Uncertainty Map, for each class (Necrotic core, Edema, and Enhancing Tumor), row-wise respectively, showing 5 slices [0,31,63,95,127], column-wise, for the axial plane, slice 95 shows the best tumor view, brighter regions indicate high variance, where the model's predictions fluctuate across runs, reflecting greater uncertainty. This is most visible near tumor boundaries, where tissue transitions are naturally ambiguous.

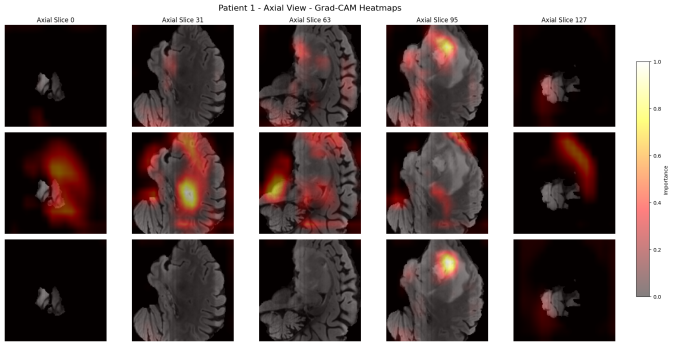

Fig. 8. Grad-CAM Heatmap, for each class (Necrotic core, Edema, Tumor, and Enhancing Tumor), row-wise respectively, showing 5 slices [0,31,63,95,127], column-wise, for the axial plane, slice 95 shows the best tumor view, highlights the regions most important for the model's prediction by weighting feature maps with their gradients. Brighter areas in the heatmap show regions that contributed more strongly to the chosen class.

class. Boundaries are ambiguous; the transition between different tissue types or tumor subregions in medical images is rarely perfectly sharp, and there is often a gradient of intensity at the edges. This indicates that the model is appropriately expressing higher uncertainty, challenging to segment.

- Grad-CAM (Gradient-weighted Class Activation Mapping) is a popular method for visualising neural network decision-making by creating a class-specific heatmap based on the gradient signal flowing into convolutional layers (Selvaraju et al., 2017). In medical imaging, Grad-CAM helps pinpoint regions of interest, such as tumor locations to increase interpretability and clinicians' trust (Suara et al., 2023).
  For 3D convolutional neural networks such as the 3D U-Net, the Grad-CAM formulation is extended by averaging gradients over the depth, height, and width of the feature maps, rather than just height and width as in the 2D case. The importance weights α_k^c for each feature map k and class c are defined as:

$$\alpha_k^c = \frac{1}{D' \cdot H' \cdot W'} \sum_{d=1}^{D'} \sum_{h=1}^{H'} \sum_{w=1}^{W'} \frac{\partial y^c}{\partial A_{d,h,w}^k}$$

  where D', H', and W' are the spatial dimensions of the 3D feature map A^k.
  Finally, the Grad-CAM heatmap for class c is computed as a weighted combination of the feature maps, followed by a ReLU operation:

$$L_c^{\text{Grad-CAM}} = \text{ReLU}\left(\sum_k \alpha_k^c A^k\right)$$

  In Figure 8, it works by using the gradients of the target class score concerning the feature maps of a chosen convolutional layer. These gradients indicate the importance of each feature map for that class, by showing the MRI scans of the modalities row-wise, T1ce, FLAIR, T2, with 5 slices [0,31,63,95,127]. The gradients are then used as weights to compute a weighted sum of the feature maps. This weighted sum is passed through a ReLU activation to highlight positive contributions, resulting in a heatmap. This heatmap is then resized to the original image dimensions. A higher intensity value in the Grad-CAM heatmap overlaid on the original image indicates that the corresponding region in the input image was more important for the model's prediction of the target class you selected for Grad-CAM. It visually shows the spatial regions that "activate" the network for a particular class.

V. Results

A. Evaluation Measures

An essential part of training any deep learning model is choosing the right loss function, which guides the model in learning. In medical image segmentation, tasks often face the challenge of class imbalance, where tumor regions are much smaller compared to the background. To address this, a combination of Dice Loss and Focal Loss was used. Dice Loss measures the overlap between predicted and actual regions, making it particularly effective for segmentation tasks where the area of interest is small. Focal Loss, on the other hand, reduces the influence of easy examples and focuses more on difficult or misclassified cases, ensuring the model learns from challenging regions. Combining these two losses helps balance precision and robustness in training (Milletari et al., 2016; Lin et al., 2017). To optimize the model's learning, the AdamW optimizer was applied. AdamW is an improved version of the standard Adam algorithm, designed to reduce overfitting through regularisation and to provide more stable convergence. It adapts the learning rate during training, while a weight decay mechanism further improves generlaisation. This makes AdamW particularly effective for medical imaging tasks where stability and accuracy are critical (Loshchilov & Hutter, 2019).

B. Implementation

1) Computing Software Requirements: To conduct this project, the BraTS 2021 dataset was obtained using IBM Aspera Connect for efficient high-speed transfer. The experiments were carried out in a Python (v3.8+) environment, either locally or using cloud platforms such as Google Colab. The segmentation model was implemented in PyTorch (v2.0+), a widely used deep learning framework. Supporting libraries were employed for data handling, numerical computation, and visualisation. Key tools include NumPy for numerical operations, Nibabel for reading medical imaging files in NIfTI format, Scikit-image for implementing the Marching Cubes algorithm, and Plotly for interactive 3D visualisation. For efficient training of the 3D U-Net, access to an NVIDIA GPU (A100) was required, which was available as a configuration option in Google Colab.

2) Training strategy: The training process was designed to optimise the model while continuously monitoring its performance. Two main evaluation metrics were used: Intersection over Union (IoU), which measures the overlap between predicted ground truth tumor regions (V7 Labs, 2023), and Dice score, a standard measure in medical image segmentation that evaluates similarity between predicted and actual regions (Milletari et al., 2016). In addition, classification accuracy was calculated to assess how well the model assigned the correct label to each voxel. During the training, batches of MRI volumes and their corresponding labels were passed through the model. The predictions were compared with the ground truth using the combined Dice and Focal loss. The model weights were updated through backpropagation, and performance metrics (IoU, Dice score, loss) were recorded across each epoch. To capture model uncertainty, Monte Carlo Dropout was applied at inference. This involved performing multiple forward passes with dropout layers active, generating slightly different predictions each time. The mean probability across was used as the final prediction, while the variance provided an estimate of prediction uncertainty. (Gal & Ghahramani, 2016). Validation followed a similar process but without updating the model's parameters. Instead, it measured performance on unseen data to assess generalisation. Metrics such as pre-class Dice, IoU, and Monte Carlo accuracy were computed and compared to training results. The overall training loop was run for multiple epochs with a cosine annealing learning rate scheduler, which gradually adjusts the learning rate following a cosine curve to improve the convergence (Loshchilov & Hutter, 2017). Performance metrics were logged and visualised throughout, and the best-performing model, based on validation loss, was saved as a checkpoint for later evaluation.

C. Segmentation Results

The 3D U-Net model was trained for 50 epochs to evaluate its performance on the brain tumor segmentation task. The training process and model performance were monitored using the Dice score, the intersection over union (IoU), and a combined loss function, with the results evaluated in the test set.

TABLE I
TEST SET EVALUATION RESULTS, WHERE CLASS 0 IS BACKGROUND, CLASS 1 IS NECROTIC CORE, CLASS 2 IS EDEMA, AND CLASS 3 IS ENHANCING TUMOR.

Metric	Class	Value
Loss	-	0.1492
Monte Carlo Accuracy	-	0.8652
4*Test IoU per class	Class 0 (Background)	0.7203
	Class 1 (Edema)	0.7702
	Class 2 (Non-enh. Tumor)	0.7785
	Class 3 (Enh. Tumor)	0.7539
4*Test Dice per class	Class 0 (Background)	0.7986
	Class 1 (Edema)	0.8443
	Class 2 (Non-enh. Tumor)	0.8501
	Class 3 (Enh. Tumor)	0.8248

The proposed 3D U-Net model demonstrates robust performance on the held-out test set, achieving a mean Dice coefficient of 0.8394 across all tumor sub-regions and an overall Monte Carlo accuracy of 86.52 percent. The model exhibits particularly strong segmentation capability for the necrotic core region (Dice:0.8443, IoU:0.7702) and edema (Dice:0.8501, IoU:0.7785), while maintaining competitive performance for the enhancing tumor region (Dice: 0.8248, IoU: 0.7539). The relatively low test loss of 0.1492 indicates good model convergence without overfitting, as evidenced by the minimal gap between validation and test performance metrics. These results are particularly noteworthy considering the model was trained for only 50 epochs due to computational constraints, suggesting that the architecture effectively learned discriminative features for multi-class brain tumor segmentation. The balanced performance across all tumor subregions demonstrates the model's ability to distinguish between complex tissue boundaries, which is crucial for clinical applications.

In Figure 9, Purple: background, Blue: necrotic core, Green: edema, Yellow: enhancing tumor The MRI scans show the modalities row-wise, T1-ce, FLAIR, T2, ground truth masks, and predicted masks.R, T2, with 5 slices [0,31,63,95,127] Slice 95 shows the best tumor view for an X-Y plane. The second-to-last row is the ground truth (segmentation mask), and the last row shows the predicted image (predicted mask).

Figure 9 illustrates the segmentation results, where purple represents the background, blue denotes edema, green corresponds to the edema, and yellow highlights the enhancing tumor. The MRI modalities are displayed row-wise as T1-ce, FLAIR, T2, each shown across five representative slices [0,31,63,95,127]. Among these, slice 95 provides the clearest visualisation of the tumor in the X-Y plane (top to bottom view, also known as axial view). The second-to-last row presents the ground truth segmentation masks, while the final row displays the model's predicted mask.

VI. DISCUSSION

A. Limitations

One of the main limitations encountered in this study was the computational resources available on Google Colab

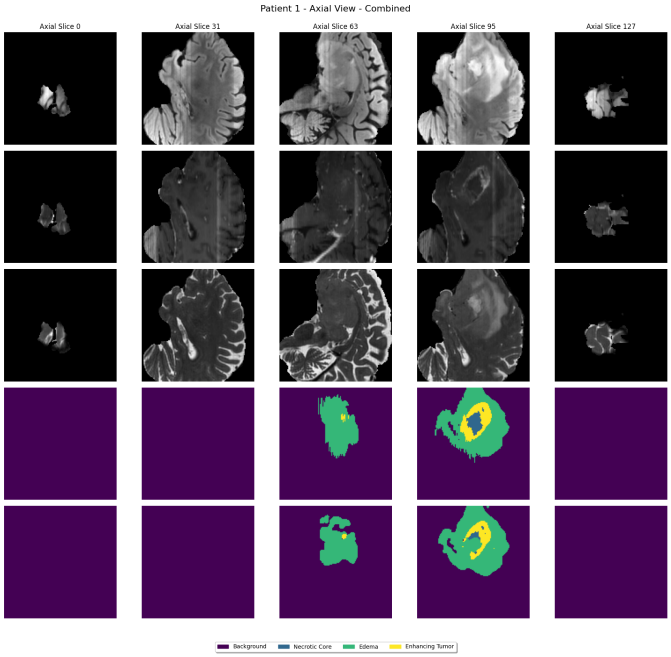

Fig. 9. Example of tumor segmentation on axial slices [0,31,63,95,127]. Rows show MRI modalities (T1ce, FLAIR, T2), ground truth masks, and predicted masks. Colour representation- purple: background; blue: necrotic core; green: edema; yellow: enhancing tumor

Pro. While Colab provides access to high-performance GPUs such as the NVIDIA A100, usage is restricted by session timeouts, memory limits, and overall compute quotas. As a result, training could not be extended beyond 50 epochs without incurring additional costs for higher-tier services. This restricted the ability to fully explore the model's convergence behaviour, as longer training or more extensive hyperparameter tuning may have further improved segmentation performance. Furthermore, the reliance on a single platform limited the opportunity to experiment with larger datasets, more complex architectures, or ensemble methods that typically require longer training times and greater computational power.

Despite demonstrating good segmentation performance, this study has several other limitations as well. First, reliance on the BraTS 2021 dataset restricts the model's generalisability to real-world hospital data with varied quality and scanner settings. Second, the use of a standard 3D U-Net architecture, while effective, may not capture the full complexity of tumor heterogeneity compared to more advanced architectures. Additionally, evaluations were limited to overlap-based metrics without external validation, which may not fully reflect clinical significance. From a practical perspective, while visualisation tools such as Grad-CAM and uncertainty maps provide interpretability, their usability by clinicians has not been fully formally assessed.

B. Future Work

To address the computational limitations faced during this project, future work could involve the use of more powerful and scalable computing infrastructure, such a high-performance computing clusters or commercial cloud platforms with extended GPU quotas. This would make it possible to train models for a larger number of epochs (e.g., 500-1000) and explore a more robust architecture. Additionally, the availability of stronger hardware would support experimentation with larger datasets and advanced data augmentation strategies, improving the model's ability to generalise to unseen clinical data, which results in better visualization outputs and more confidence in the model.

A key area for improving the web application would be the addition of a dynamic file upload feature. This would enable users, such as clinicians or researchers, to upload their own NIfTI files directly through the interface and generate a 3D visualisation of their specific data, rather than being restricted to the preloaded dataset stored on the server.

VII. Conclusion

This dissertation successfully developed and evaluated a deep learning framework for the automated segmentation of gliomas and their constituent subregions from multimodal MRI scans, utilizing the BraTS 2021 dataset. The primary objective was to leverage the full spatial context of volumetric medical data, for which a 3D U-Net architecture was implemented and trained. The model demonstrated strong performance, effectively learning the complex hierarchical features necessary to distinguish between healthy tissue and pathological subregions, namely the necrotic core, peritumoral edema, and the enhancing tumor.

The methodology was grounded in established best practices, including comprehensive data preprocessing, a combined Dice and Focal Loss function to mitigate class imbalance, and the AdamW optimizer for robust convergence. Quantitative evaluation, based on per-class Dice and Intersection over Union (IoU) metrics, confirmed the model's high accuracy and its ability to generalize from the training data to unseen validation cases.

Beyond segmentation accuracy, this work placed a significant emphasis on model interpretability and reliability. The implementation of Grad-CAM provided crucial visual evidence of the model's decision-making process, generating heatmaps that highlighted the specific anatomical regions influencing its prediction for each voxel measure of the model's confidence, revealing its certainty in core tumor regions and its relative uncertainty at ambiguous boundaries. Finally, Monte Carlo Dropout was employed to quantify epistemic uncertainty, producing variance maps that effectively captured where the model was 'guessing', particularly at the subtle edges between different tissue types.

In conclusion, this project demonstrates that a 3D U-Net, coupled with advanced interpretability techniques, serves as a powerful and transparent tool for automated brain tumor segmentation. The ability to not only produce accurate segmentation masks but also to provide insights into the model's confidence and reasoning is a critical step towards building trust and facilitating the adoption of such AI-driven tools in a clinical setting. The visualization and analyses performed

in this work lay the groundwork for a more comprehensive understanding of tumor morphology, which can ultimately aid clinicians in diagnosis, treatment planning, and patient monitoring. The application leverages the Marching Cubes algorithm to generate interactive 3D surface models of the brain and constituent tumor subregions. This provides a powerful and accessible platform for the qualitative analysis and intuitive understanding of complex volumetric medical data.

VIII. References

1) Teng, Q., Liu, Z., Song, Y., Han, K. & Lu, Y. (2022) 'A survey on the interpretability of deep learning in medical diagnosis', Multimedia Systems, 28, pp. 2335–2355.
2) Salahuddin, Z., Woodruff, H. C., Chatterjee, A. & Lambin, P. (2021) 'Transparency of deep neural networks for medical image analysis: a review of interpretability methods', Computers in Biology and Medicine, 140, 105111.
3) Hanif, A. M. et al. (2021) 'Applications of interpretability in deep learning models for ophthalmology'.
4) Vellido,A. (2022) 'The importance of interpretability and visualization in machine learning for applications in medicine and healthcare', Neural Computing and Applications, 32, pp. 18069–18083.
5) Alam,K.N. et al. (2025) 'Attribution-based explainability in medical imaging', Electronics, 14(15), 3024.
6) Huff,D.T., Weisman,.J. & Jeraj,R. (2021) 'Interpretation and visualization techniques for deep learning models in medical imaging', Medical Physics, University of Wisconsin–Madison.
7) Lambert,B., Forbes,F., Tucholka,A., Doyle,S., Dehaene,H. & Dojat,M. (2022) 'Trustworthy clinical AI solutions: a unified review of uncertainty quantification in deep learning models for medical image analysis', arXiv preprint.
8) Ronneberger, O., Fischer, P., & Brox, T. (2015). U-Net: Convolutional Networks for Biomedical Image Segmentation. In Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015 (pp. 234-241). Springer.
9) Çiçek, Ö., Abdulkadir, A., Lienkamp, S. S., Brox, T., & Ronneberger, O. (2016). 3D U-Net: Learning Dense Volumetric Segmentation from Sparse Annotation. In Medical Image Computing and Computer-Assisted Intervention – MICCAI 2016 (pp. 424-432). Springer.
10) Milletari, F., Navab, N., & Ahmadi, S. A. (2016). V-Net: Fully Convolutional Neural Networks for Volumetric Medical Image Segmentation. In 2016 Fourth International Conference on 3D Vision (3DV) (pp. 565-571). IEEE.
11) Isensee, F., Jaeger, P. F., Kohl, S. A. A., Petersen, J., & Maier-Hein, K. H. (2021). nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. Nature Methods, 18(2), 203-211.
12) Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization. In Proceedings of the IEEE International Conference on Computer Vision (ICCV) (pp. 618-626).
13) Gal, Y., & Ghahramani, Z. (2016). Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning. In Proceedings of the 33rd International Conference on Machine Learning (ICML) (pp. 1050-1059)
14) Hähn, D., Rannou, N., Grant, P.E. and Pienaar, R. (2012) Slice:Drop: A platform for online 3D visualization of medical image data.
15) Menze, B.H. et al. (2015) 'The Multimodal Brain Tumor Image Segmentation Benchmark (BRATS)', IEEE Transactions on Medical Imaging, 34(10), pp. 1993–2024.
16) Bakas, S. et al. (2018) 'Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the BRATS challenge', arXiv preprint, arXiv:1811.02629.
17) Bakas, S. et al. (2017) 'Advancing the cancer genome atlas glioma MRI collections with expert segmentation labels and radiomic features', Scientific Data, 4, 170117.
18) Lorensen, W.E. & Cline, H.E. (1987) 'Marching Cubes: A high resolution 3D surface construction algorithm', ACM SIGGRAPH Computer Graphics, 21(4), pp. 163–169.
19) Number Analytics (2025) Ultimate guide to softmax in neural networks and deep learning. Available at: https://www.numberanalytics.com/blog/ultimate-guide-softmax-neural-networks-deep-learning
20) Mackinlay, D. (2024) Softmax. Available at: https://danmackinlay.name/notebook/softmax.html
21) Kendall, A. & Gal, Y. (2017) 'What uncertainties do we need in Bayesian deep learning for computer vision?', Advances in Neural Information Processing Systems (NeurIPS), 30, pp. 5574–5584.
22) Abdar, M. et al. (2021) 'A review of uncertainty quantification in deep learning: Techniques, applications and challenges', Information Fusion, 76, pp. 243–297. doi: 10.1016/j.inffus.2021.05.008
23) Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D. & Batra, D. (2017) 'Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization', in Proceedings of the IEEE International Conference on Computer Vision (ICCV), pp.618–626.
24) Suara, S., Jha, A., Sinha, P. & Sekh, A.A. (2023) 'Is Grad-CAM explainable in medical images?', ArXiv preprint.
25) Lin, T.Y., Goyal, P., Girshick, R., He, K. & Dollár, P. (2017) 'Focal Loss for Dense Object Detection', Proceedings of the IEEE International Conference on Computer Vision (ICCV), pp. 2980–2988.
26) Loshchilov, I. & Hutter, F. (2019) 'Decoupled weight decay regularization', International Conference on Learning Representations (ICLR).

27) V7 Labs (2023) Intersection over Union (IoU): A complete guide. Available at: https://www.v7labs.com/blog/intersection-over-union-guide
28) Loshchilov, I. & Hutter, F. (2017) 'SGDR: Stochastic gradient descent with warm restarts', International Conference on Learning Representations (ICLR).
29) Plotly Technologies Inc. (2015) Collaborative data science. Montréal: Plotly Technologies Inc. Available at: https://plot.ly
30) Datature (2025) A comprehensive guide to 3D models for medical image segmentation. Available at: https://datature.com/blog/a-comprehensive-guide-to-3d-models-for-medical-image-segmentation