A dissertation submitted to the **University of Greenwich**
in partial fulfilment of the requirements for the Degree of

Master of Science
in
Data Science

Brain Tumor Detection: Classification and Survival Prediction

Name: Anushka Pradeep Kadam

Student ID: 001311625

Supervisor : Dr. Mohammad Al-Antary

Submission date : 6th September 2024

Word Count : 14,930

ABSTRACT

This study provides a comprehensive exploration into the application of advanced deep learning and machine learning techniques for brain tumor classification, segmentation, and survival prediction using MRI data. The research employed transfer learning models—InceptionV3, MobileNetV2, ResNet152V2, and VGG19—with pre-trained weights from the ImageNet dataset to address the classification task.

Among these, the InceptionV3 model exhibited the highest accuracy, achieving 92.83%, with ensemble methods, particularly the weighted average ensemble, slightly surpassing it at 92.98%. This finding underscored the potential of model combination for enhanced accuracy. In contrast, ResNet152V2, despite its considerable depth and complexity, underperformed with an accuracy of 68.95%, highlighting the critical importance of selecting models that are well-suited to the dataset's specific characteristics. MobileNetV2, noted for its efficiency, achieved a commendable balance between accuracy and speed, registering 90.54%, thereby rendering it suitable for real-time applications.

For segmentation, the study developed and optimized both a customized 3D U-Net and a 3D Attention U-Net for processing volumetric MRI data. The 3D U-Net achieved a mean Intersection over Union (IoU) score of 74.54% on validation data, while the 3D Attention U-Net, enhanced with attention mechanisms to focus on critical regions, attained a slightly lower mean IoU of 69.19%. The attention mechanisms proved particularly advantageous in improving the segmentation of complex and smaller tumor regions, such as the enhancing tumor and edema, despite some fluctuations observed during training.

Survival prediction was approached through traditional machine learning models, including Random Forest, Gradient Boosting, Support Vector Machines (SVM), and a Voting Classifier ensemble. The Random Forest model achieved the highest training accuracy at 71%, though its validation accuracy declined to 56%, indicating potential overfitting. The Voting Classifier demonstrated a more balanced performance with a validation accuracy of 56%, suggesting that ensemble methods could mitigate some overfitting issues observed in individual models. However, the overall performance was constrained by the simplicity of the dataset, which included only age, tumor class weights, and survival days as features.

ACKNOWLEDGEMENTS

I would especially like to thank Dr. Mohammad Al-Antary for agreeing to be my supervisor and for his consistent advice, feedback, guidance, and support throughout the lifecycle of this MSc Data Science project.

I want to thank both Dr. Mohammad Al-Antary and Dr. Pushparajah Rajaguru for agreeing to have the project demonstration on the schedule day.

A special thanks to all my friends who made this experience another memorable chapter of my life, and for supporting me throughout this academic year.

TABLE OF CONTENT

INDEX OF FIGURES

INDEX OF TABLES

CHAPTER 1
1. INTRODUCTION

1.1 Overview

This dissertation explores the integration of medical imaging and artificial intelligence, focusing specifically on brain tumor detection, classification, and survival prediction using MRI data. The rapid progress in deep learning and machine learning has enabled the development of more accurate and efficient diagnostic tools in neuro-oncology. This study leverages advanced convolutional neural networks (CNNs) and ensemble techniques to improve the precision of brain tumor classification and segmentation. By employing models such as InceptionV3, MobileNetV2, ResNet152V2, and VGG19, alongside a 3D U-Net and 3D Attention U-Net architectures for segmentation, the research seeks to align clinical requirements with technological advancements. For survival prediction, traditional machine learning models were utilized, with Random Forest achieving the highest training accuracy, albeit with signs of overfitting. The use of ensemble methods offered a balanced performance, though the study acknowledges the limitations posed by the simplicity of the dataset.

1.2 Background

Brain tumors are among the most complex and lethal types of cancer, profoundly affecting both the structural and functional aspects of the central nervous system (CNS) (DeAngelis, 2001). They are a diverse group of neoplasms that originate in the brain or CNS, categorized based on their origin and malignancy. Tumors are either primary, developing within the brain, or secondary (metastatic), spreading to the brain from other parts of the body (The ASCO foundation, 2022).

According to the World Health Organization (WHO), tumors are classified by malignancy into grades I to IV, with increasing aggressiveness (Kleihues, Burger, & Scheithauer, 1993). High-Grade Gliomas (HGG), including grades III and IV, necessitate immediate treatment due to their rapid progression and potential to cause death within two years. Conversely, Low-Grade Gliomas (LGG) are benign, growing slowly and allowing patients several years of life expectancy. Brain tumors can be malignant (cancerous), typically high grade (grade 3 or 4),

and either originate in the brain (primary tumors) or spread to the brain from other areas (secondary tumors) can been seen in **Figure 1**. These tumors are more likely to recur after treatment. Alternatively, benign (non-cancerous) tumors, which are low grade (grade 1 or 2), grow slowly and are less likely to return post-treatment (Louis, 2007) (Jovčevska I, 2013).

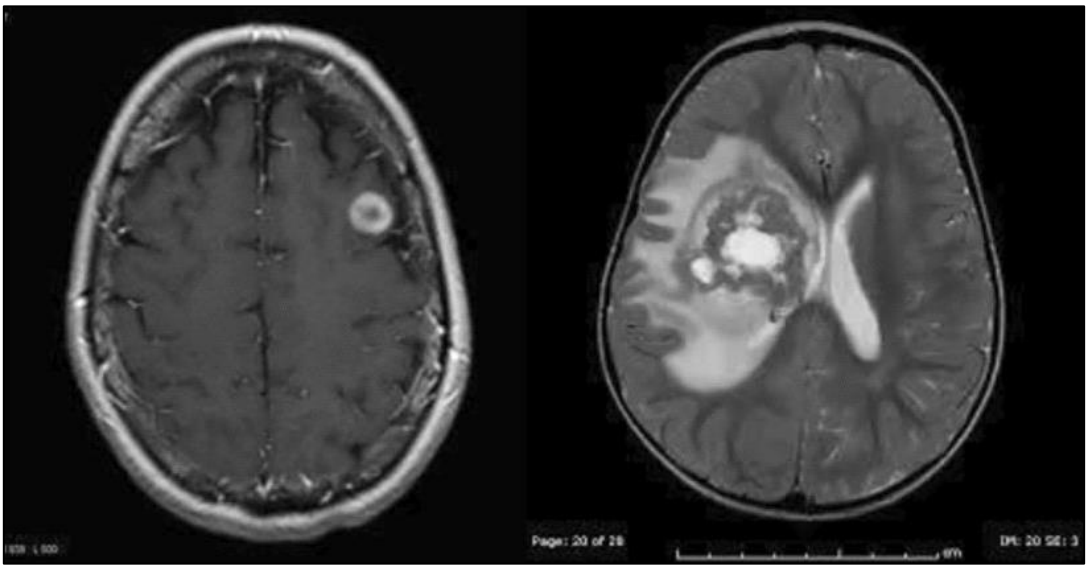


Figure 1. MRI scans of Brain Tumor (LEFT- BENIGN & RIGHT-MALIGANT)

Source- *(Paul, 2021)*

Gliomas, particularly glioblastomas—a subtype of gliomas—are often malignant, presenting significant treatment challenges due to their aggressive nature and diffuse infiltration into surrounding brain tissue. Benign tumors, such as meningiomas, generally grow slowly and are less likely to invade adjacent brain tissues (Louis DN, 2016) (Rehman A. N., 2020). However, malignant tumors, including glioblastomas, exhibit rapid and aggressive growth, underscoring the importance of early detection and precise classification for effective treatment and improved survival rates.

Importance of Early Detection and Accurate Classification

Early detection and accurate classification of brain tumors are crucial for several reasons. Firstly, they enable timely and appropriate therapeutic interventions, significantly improving patient outcomes. For instance, benign tumors like meningiomas can often be managed with surgical resection alone, while malignant tumors such as glioblastomas require a combination of surgery, radiation therapy, and chemotherapy. Misclassification can lead to suboptimal

treatment plans, potentially compromising patient outcomes. Secondly, precise classification helps in tailoring personalized treatment plans, enhancing the effectiveness of therapies, and reducing the risk of adverse effects. Lastly, accurate prognostic assessments provide valuable information to patients and their families, aiding in decision-making and improving overall care quality.

Traditional brain tumor diagnosis methods include clinical evaluations, neuroimaging techniques such as Magnetic Resonance Imaging (MRI) and Computed Tomography (CT) scans, and histopathological examination of biopsy samples. While effective, these methods have limitations. For instance, biopsy procedures are invasive and carry risks, and histopathological analysis can be subject to interobserver variability. Moreover, manual interpretation of medical images is time-consuming and prone to errors, particularly in distinguishing subtle differences in tumor types and grades. (al, 2021).

Advancements in Medical Imaging and Machine Learning

Recent advancements in medical imaging and computational methods have significantly improved brain tumor detection and classification. MRI is widely regarded as the optimal method for detecting brain tumors due to its superior soft tissue contrast (Liang Z.-P. a., 2000). The use of Gadolinium-enhanced MRI sequences further enhances image clarity, improving tumor detection and diagnosis (Bauer, 2013). However, manual tumor diagnosis using MRI is challenging due to the vast amount of data involved (Bankman, 2008), driving the development of automated or semi-automated brain tumor segmentation techniques.

These automated segmentation techniques can be categorized into basic, generative, and discriminative methods (Menze, 2015) (Agravat, Deep Learning for Automated Brain Tumor Segmentation in MRI Images, 2018). With the advent of deep learning, advanced methods leveraging Convolutional Neural Networks (CNNs) have become the standard for tumor segmentation (Agravat, Brain Tumor Segmentation, 2016). CNNs automatically learn hierarchical feature representations from raw imaging data, making them particularly effective for complex image analysis tasks, such as brain tumor classification.

In parallel, machine learning, a subset of artificial intelligence, has advanced significantly, with sophisticated algorithms capable of analysing complex datasets and extracting meaningful patterns. In medical imaging, machine learning algorithms can be trained to recognize and

classify different types of brain tumors based on imaging features, offering the potential for highly accurate, automated diagnostic tools (Kshatri, 2023).

1.3 Objectives of the Study

This dissertation aims to harness the power of ML and DL to develop robust models for brain tumor detection, classification, and survival prediction. The specific objectives of this study are as follows:

Comparative Analysis of Transfer Learning Models for Brain Tumor Classification:

This study aims to evaluate and compare the performance of four transfer learning models—InceptionV3, MobileNetV2, ResNet152V2, and VGG19—in classifying brain tumors. By leveraging pre-trained models on the ImageNet dataset, the objective is to identify which architecture best balances accuracy, efficiency, and computational cost in the context of medical imaging, particularly for brain tumor classification.

Accurate Tumor Segmentation Using 3D U-Net & 3D Attention U-Net:

Another critical objective is the implementation and fine-tuning of a 3D U-Net model, specifically tailored for the segmentation of 3D MRI volumes. This involves accurately delineating tumor boundaries, which is essential for treatment planning and monitoring. The study will explore advanced techniques within the 3D U-Net architecture, such as skip connections and upsampling, to retain critical spatial information and improve segmentation accuracy. This objective seeks to push the boundaries of what is achievable in automated tumor segmentation, moving towards more reliable and precise outcomes.

Survival Prediction Using Clinical and Imaging Data:

The final objective is to predict patient survival outcomes by developing a model that integrates radiomic features extracted from MRI images with clinical data, such as age and tumor characteristics.

CHAPTER 2

2. LITERATURE REVIEW

2.1 Overview

The advancement of methodologies for brain tumor detection, classification, segmentation and survival prediction has been a pivotal area of research in recent years. These advancements are crucial, given the complexity and lethality of brain tumors, which significantly impact patients' quality of life and survival rates. Traditional diagnostic techniques, although effective, often have limitations that necessitate the development of more sophisticated approaches. This literature review explores the significant contributions of various researchers in this domain, highlighting their methodologies, findings, and the potential impact on medical diagnosis and treatment.

2.2 Similar research and tools

2.2.1 Brain Tumor Classification

One notable contribution to the exploration of Convolutional Neural Network (CNN) architectures is the minimalist approach proposed by Abiwinanda et al. They investigated the use of a basic Convolutional Neural Network (CNN) for classifying Glioma, Meningioma, and Pituitary tumors using a dataset of 3064 T-1 weighted CE-MRI images. The model achieved a training accuracy of 98.51% and a validation accuracy of 84.19%, demonstrating that even a simple CNN architecture can effectively classify brain tumors without the need for complex segmentation techniques. (Abiwinanda, 2019).

Sajjad et al. concentrated on the classification of glioma, meningioma, and pituitary tumors. Utilizing GoogleNet for feature extraction from MRI images and employing a rigorous fivefold cross-validation technique, they achieved an impressive model accuracy rate of 98% (Sajjad M, 2019). Similarly, Lakshmi et al. proposed a two-step Computer-Aided System (CAD) designed to automatically detect and categorize brain tumors as either malignant or benign (Devasena, 2013). This system employed feature extraction techniques such as Principal Component

Analysis (PCA) and Discrete Wavelet Transform (DWT), followed by Support Vector Machine (SVM) classification.

Khan et al. introduced a novel deep learning approach for classifying brain tumors into cancerous and non-cancerous categories using real brain MRI scans with data augmentation. Their method incorporated edge detection and feature extraction via a simple CNN model, resulting in an 89% classification accuracy (Khan HA, 2020). Further advancing the field, Kabir Anaraki et al. combined CNN and genetic algorithms (GA) for the non-invasive classification of glioma grades, achieving accuracies of 90.9% for three glioma grades and 94.2% for distinguishing glioma, meningioma, and pituitary tumor types (Kabir Anaraki, 2019).

To address the challenge of limited data availability, Ertosun et al. developed a robust deep learning pipeline employing an ensemble of CNNs (Ertosun MG, 2015). This method achieved notable accuracies of 96% for distinguishing high-grade gliomas (HGG) from low-grade gliomas (LGG) and 71% for classifying LGG Grade I versus LGG Grade II. Tahir et al. explored various methods to refine classification accuracy, emphasizing the importance of integrating diverse data types and employing a blend of preprocessing techniques. Their suggested model achieved an accuracy of 86% (Tahir, 2019).

Sachdeva et al. introduced an innovative method using Probabilistic Neural Networks (PNN), integrating data and image processing techniques. Their approach involved feature extraction via PCA, followed by feeding the extracted features into the PNN for classification purposes (Sachdeva, 2013). Rehman et al. utilized three pre-trained CNN architectures: AlexNet, GoogleNet, and VGG16, for brain tumor classification. Employing transfer learning techniques, the VGG16 model emerged as the most effective, achieving a classification accuracy of 98.69% (Rehman A. N., 2020).

Mehrotra et al. revolutionized brain tumor classification using deep learning and transfer learning techniques. They analyzed T1-weighted MRI images to differentiate between malignant and benign tumors, exploring various state-of-the-art CNN models. The pre-trained AlexNet CNN model, leveraging transfer learning, achieved an astounding accuracy rate of 99.04%, signaling a significant breakthrough in the field (Mehrotra R, 2020).

Casamitjana et al. introduced a machine learning framework for the precise detection and classification of brain tumors in MRI data, utilizing techniques like PCA and DWT (Casamitjana, 2016). Goswami et al. pioneered the "Hybrid Abdominal Detection Algorithm" for identifying irregularities within the body using MRI images (Goswami, 2013). Saltz et al. devised a hierarchical approach to segment brain tumors, employing a specialized CNN architecture and evaluating performance using metrics such as the Dice Score Coefficient, Positive Predictive Value, and Sensitivity (Saltz J, 2018).

Afshar et al. presented CapsNet, an innovative CNN architecture tailored for brain tumor classification (Afshar, 2018). This approach capitalizes on the spatial relationships between tumors and surrounding tissues, achieving exceptional accuracy rates of 86.56% for segmented tumors and 72.13% for unprocessed brain images. Jude Hemanth et al. introduced novel adaptations of Artificial Neural Networks, namely Multi-Class Probabilistic Neural Network (MCPN) and Multi-Kernel Neural Network (MKNN), to overcome convergence time challenges and enhance accuracy in detecting brain abnormalities (Jude Hemanth, 2014).

2.2.2 Segmentation and Survival Prediction Techniques

With the advent of deep learning, advanced methods leveraging Convolutional Neural Networks (CNNs) have become the standard for tumor segmentation (Agravat, Brain Tumor Segmentation, 2016) . These approaches often involve further segmenting the tumor into substructures, such as necrosis, enhancing tumor, and edema, which are critical for accurate survival prediction. Tumor size and the extent of its substructures are key factors in predicting overall survival (OS).

For instance, a 3D U-Net-based model was employed for tumor segmentation, with radiomics features extracted from segmentation masks used for OS prediction. A Random Forest Regressor (RFR) with 1000 trees, combined with an ensemble of small multilayer perceptrons (MLPs), achieved an accuracy of 52.6% on the test dataset, with a Spearman correlation coefficient of 0.496 (Isensee, 2017).

In another study (Chato, 2017), a pre-trained AlexNet was used for tumor segmentation, and the extracted features were applied to a linear discriminant model for OS prediction, yielding an accuracy of 46% for texture features and 68.5% for histogram features on the test dataset. Furthermore, a fully automated model for segmenting Low-Grade Gliomas (LGG) and High-

Grade Gliomas (HGG) in multimodal MRIs achieved 100% accuracy in OS prediction using Support Vector Machine (SVM) algorithms on a small test set of 16 samples (Osman, 2017). Another approach utilized the Dense-Res-Inception Net (DRINet) for biomedical image segmentation, reporting Dice Similarity Coefficients (DSCs) of 83.47% for the whole tumor, 73.41% for the tumor core, and 64.98% for the enhancing tumor (Liang C. B., 2018).

Other notable methods include a Fully Convolutional Neural Network (FCNN) architecture that was applied for tumor segmentation and OS prediction using SVM classifiers (Varghese, 2017). This method incorporated Z-score normalization to correct for multi-center data variability and magnetic field inhomogeneities, achieving a 60% accuracy for OS prediction and achieving Dice scores of 0.83 for whole tumor, 0.69 for tumor core, and 0.69 for active tumor regions on the BraTS 2017 dataset. An ensemble model combining 19 variations of DeepMedic and 7 variations of 3D U-Net, utilizing features such as age, spatial, volumetric, and morphological characteristics, attained an accuracy of 70% using ground truth features and 63% using network segmentation features on a dataset of 59 patients (Kao, 2019).

In another study (Gates, 2019), the DeepMedic CNN architecture was used for tumor segmentation, with OS prediction implemented via the Cox model, resulting in DSCs of 80%, 68%, and 67% for the whole tumor, tumor core, and enhancing tumor, respectively. The accuracy of OS prediction was 44.5% for the training set and 38.2% for the test set. The PixelNet architecture (Islam, 2019) achieved 88% DSC for whole tumor segmentation and a 54.5% accuracy in OS prediction using an Artificial Neural Network (ANN) trained on mean, skewness, and tumor location features. Other approaches, such as a densely connected CNN for segmentation combined with an MLP-based regressor for OS prediction, reported a 50% accuracy on the training data (Kori, 2019). An ensemble of three convolution networks with a hybrid loss function achieved a 52.6% accuracy on the validation set (Ren, 2018), while modifications to the U-Net architecture with bottleneck and dense layers, combined with elastic net for OS prediction, resulted in a 67% accuracy on the training data (Shin, 2018).

Extended U-Net architectures (Xu, 2018) and residual U-Net implementations (Yang, 2018) have also been explored for tumor segmentation and OS prediction, achieving accuracies of 65% and 47.5%, respectively, in training data.

2.3 Conclusion

The fields of brain tumor classification, segmentation, and survival prediction have advanced significantly due to persistent research efforts worldwide. By employing cutting-edge techniques such as deep learning, transfer learning, and innovative CNN architectures, substantial improvements in diagnostic accuracy and efficiency were achieved. Diverse methodologies, from minimalist CNN designs to sophisticated ensemble models and hybrid techniques, have contributed to these advancements. However, challenges persist, particularly in managing limited data, enhancing generalization to unseen datasets, and optimizing models for clinical application.

This project aimed to address these gaps by exploring transfer learning, ensemble techniques, and advanced models like 3D U-Net and 3D Attention U-Net for multimodal images. Transfer learning, utilizing pre-trained models on large datasets, proved effective in mitigating data limitations—a prevalent issue in medical imaging. This technique was particularly valuable for brain tumor classification and segmentation, where annotated datasets are scarce but essential for training robust models. Ensemble models, which combine the outputs of multiple classifiers, demonstrated superior performance over individual models by capturing diverse data patterns and reducing overfitting.

The use of 3D U-Net and its attention-enhanced variant represented state-of-the-art approaches for accurate tumor segmentation in multimodal MRI images. These models allowed for detailed segmentation of tumor substructures—such as necrosis, edema, and enhancing tumor—providing essential features for accurate survival prediction. The integration of different MRI modalities (T1, T2, FLAIR) provided complementary information, enhancing segmentation and prediction accuracy. The selection of 3D U-Net and 3D Attention U-Net in a multimodal context leveraged this data richness, making the models more adaptable and reliable for clinical use.

Beyond segmentation, the project also emphasized survival prediction, aiming to offer actionable insights into patient outcomes. Considering the complexities of brain tumors, a combination of volumetric features from segmented tumor regions, clinical data (such as age), and advanced machine learning models (Random Forest, Gradient Boosting, and SVM) was used to develop a comprehensive approach to survival prediction.

CHAPTER 3

3. MACHINE LEARNING

3.1 Introduction to Machine Learning

Machine learning (ML) is a key branch of artificial intelligence (AI) that allows computers to learn from data and make decisions without explicit programming. Unlike traditional programming, where outcomes are determined by set instructions, ML models learn from sample data (training data) to predict or decide on new, unseen data. This adaptive learning capability makes ML particularly effective for complex, evolving problems. (IBM, 2024)

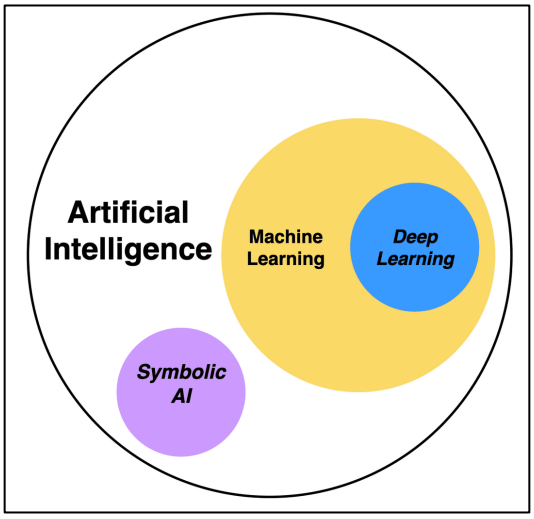


Figure 2. Relationship Between AI, ML & DL

Source- *(Amazon AWS, n.d.)*

Machine learning is broadly categorized into three types:

Supervised Learning: Involves training models on labeled datasets where each input is paired with a known output, allowing the model to learn and predict outcomes for new data. It's widely used in classification (e.g., detecting tumors in MRI scans) and regression tasks (e.g., predicting survival times) (IBM, 2024).

Unsupervised Learning: Works with unlabeled data to identify patterns and relationships. Applications include clustering similar data points (e.g., grouping patients by tumor characteristics) and reducing data complexity (e.g., dimensionality reduction) (IBM, 2024).

Reinforcement Learning: Works with unlabeled data to identify patterns and relationships. Applications include clustering similar data points (e.g., grouping patients by tumor characteristics) and reducing data complexity (e.g., dimensionality reduction) (IBM, 2024).

The convergence of ML and medical imaging has significantly enhanced diagnostic capabilities, particularly with deep learning models like Convolutional Neural Networks (CNNs). As shown in the **Figure 3**, traditional machine learning separates feature extraction and classification, while CNNs combine these steps within a single model, automating the process. This capability makes CNNs ideal for tasks such as image classification and

segmentation, crucial in medical applications including brain tumor analysis where they can directly learn complex patterns from imaging data, streamlining and improving diagnostic accuracy. (IBM, 2024).

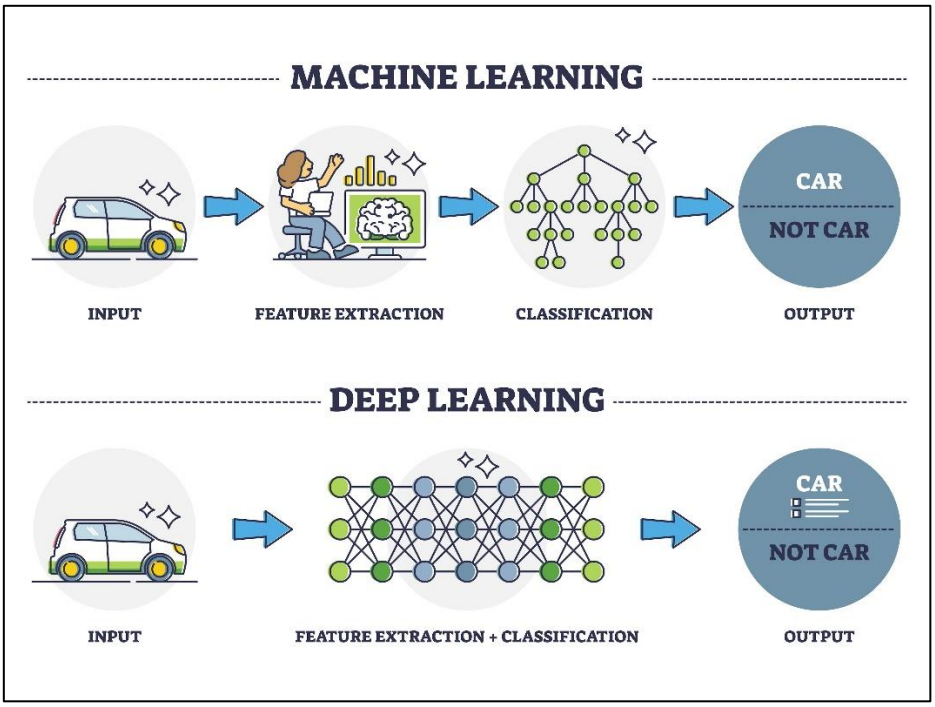


Figure 3. Machine Learning vs Deep Learning
Source- *(turing, 2024)*

3.2 Convolution Neural Networks (CNNs)

Convolutional Neural Networks (CNNs) are fundamental in the realm of image analysis, particularly for tasks like brain tumor classification and segmentation. The provide **Figure 4** illustrates the typical structure of a CNN, which includes key components such as convolutional layers, pooling layers, and fully connected layers, each playing a vital role in processing visual data.

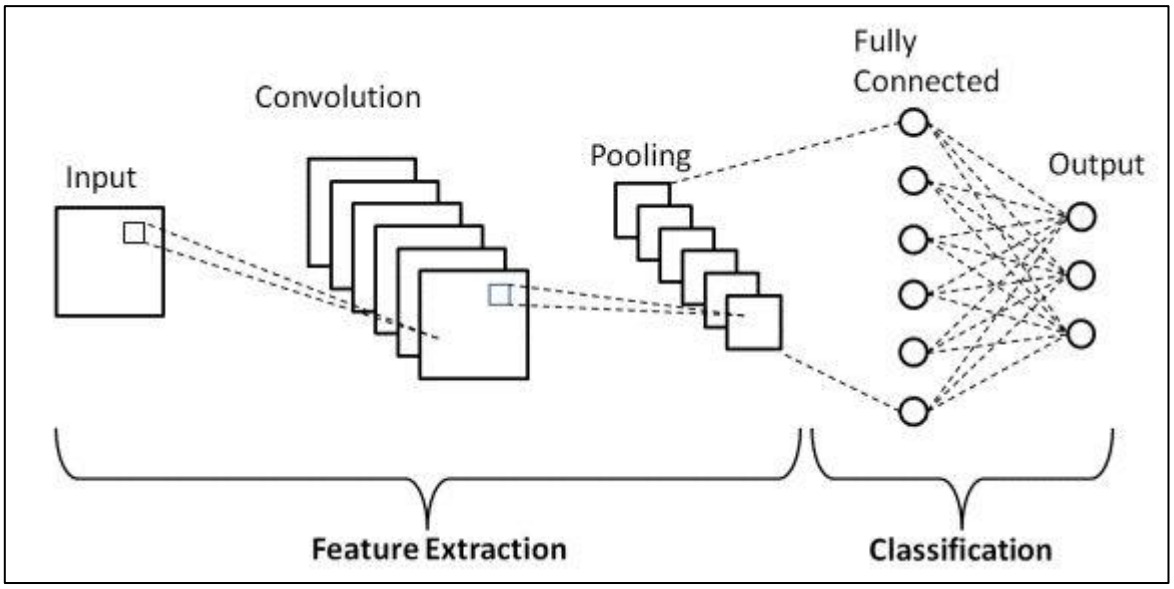


Figure 4. Basic CNN Architecture Source- *(Gurucharan, 2024)*

Convolutional layers are the backbone of CNNs, using filters to scan images and extract features like edges and textures. This enables the model to learn spatial hierarchies, crucial for distinguishing between various brain tumor types. In your models—InceptionV3, MobileNetV2, VGG19, and 3D U-Net—these layers detect intricate patterns in MRI scans. (Alzubaidi, 2021)

Pooling layers, especially Max Pooling, are used to reduce the size of the data while retaining important features. Max Pooling selects the highest value from regions of the feature map, emphasizing the most prominent features, which is vital for capturing distinct tumor characteristics. This approach is preferred over Average Pooling and Global Pooling for its ability to maintain critical information efficiently. (Gurucharan, 2024)

Fully connected layers transform the extracted features into a one-dimensional vector for classification, consolidating all learned patterns to produce final outputs like class probabilities.

Activation functions introduce non-linearity, enabling the network to learn complex patterns. ReLU (Rectified Linear Unit) is widely used in your models for its simplicity and effectiveness in preventing the vanishing gradient problem, thereby enhancing training speed. For segmentation tasks in the 3D Attention U-Net, **Leaky ReLU** is used to avoid inactive neurons by allowing a small gradient for negative inputs, improving focus on subtle regions like tumor cores and edema. (Alzubaidi, 2021)

Finally, **Softmax** is used in classification outputs to convert logits into probabilities, ideal for multi-class tasks like tumor type differentiation. This combination of convolutional layers, max

pooling, and specific activation functions ensures that your CNNs are both robust and highly effective for the challenges of medical imaging. (Alzubaidi, 2021) (Gurucharan, 2024)

3.3 Transfer Learning

Transfer learning is a machine learning technique where a pre-trained model, originally developed for a different but related task, is adapted to a new task. This approach leverages the knowledge gained from the original task to improve the performance and efficiency of the model on the new task. Transfer learning is particularly useful when there is a limited amount of labeled data available for the new task, as it allows the model to benefit from the large datasets used to train the original model. (Donges, 2024)

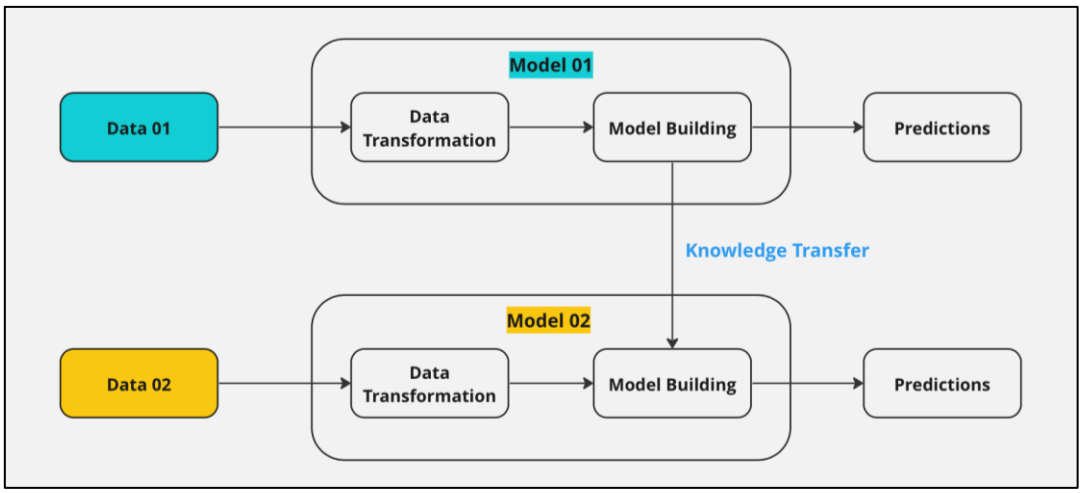


Figure 5. Tranfer Learning Flow Diagram
Source- *(botpenguin, n.d.)*

3.3.1 Concept of Transfer Learning

The fundamental idea behind transfer learning is to take advantage of the representations learned by a model on a large and diverse dataset and apply those representations to a related but different problem. This process involves several steps:

- Pre-training: The model is first trained on a large dataset, often on a general task such as image classification using datasets like ImageNet. (Donges, 2024)
- Transfer: The pre-trained model, including its learned weights and features, is transferred to the new task. The earlier layers, which capture general features such as edges and textures, are typically retained. (Donges, 2024)

- Fine-tuning: The model is then fine-tuned on the new dataset, with the later layers being adjusted to learn the specific features relevant to the new task. This step may involve freezing the weights of the earlier layers and only training the final layers or training the entire model with a lower learning rate. (Donges, 2024)

3.3.2 Application in Brain Tumor Detection

In this study, transfer learning was employed to develop an effective brain tumor detection model. Several pre-trained CNN architectures were explored, including:

- **InceptionV3 Model:**

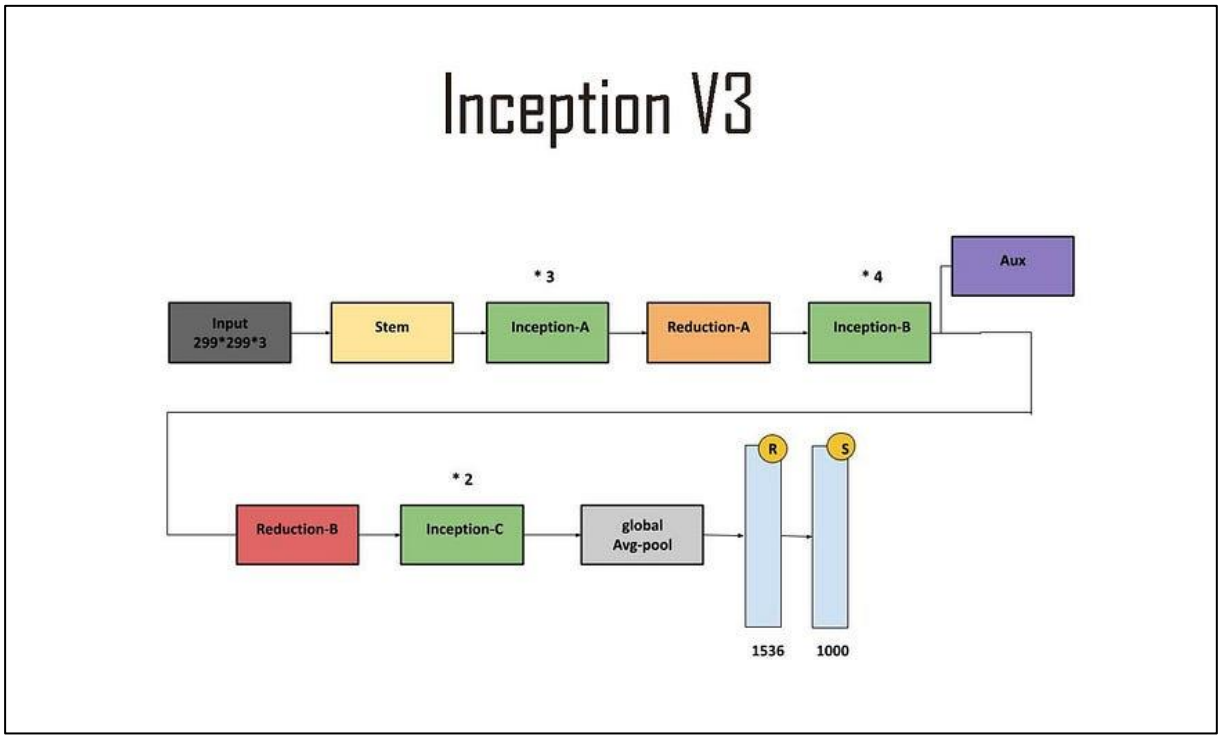


Figure 6. InceptionV3 architecture

Source- *(Brital, 2021)*

Introduced by Christian Szegedy et al. in 2016, InceptionV3 is a sophisticated evolution of the original Inception architecture. Its modular design allows the model to effectively learn from images at multiple scales. InceptionV3 employs parallel convolutional layers of varying filter sizes within Inception modules, enabling simultaneous analysis of different spatial resolutions. The architecture also incorporates Reduction modules for down sampling feature maps while preserving crucial information, and a global average pooling layer to distill feature maps into single values per map. Additionally, an auxiliary classifier improves gradient flow during

training, leading to better model convergence. This design strikes a balance between computational efficiency and high accuracy, making InceptionV3 highly effective for large-scale image classification tasks. (Brital, 2021)

- **MobilenetV2 Model:**

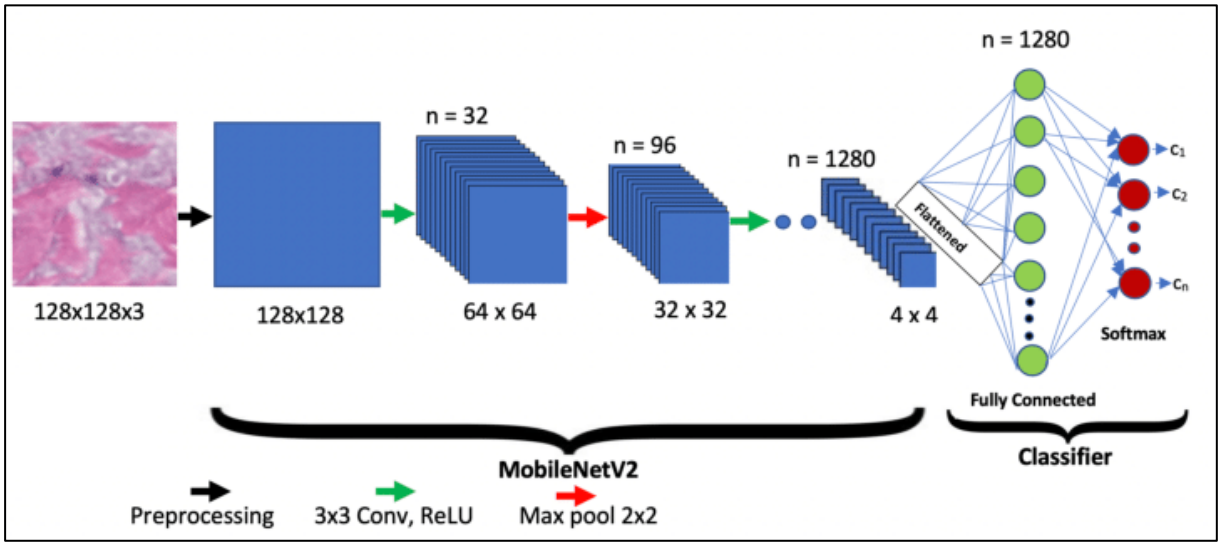


Figure 7. MobilenetV2 Architecture

Source- *(Akay, 2021)*

MobileNetV2 is a streamlined and efficient convolutional neural network architecture designed for mobile and edge devices. Introduced by Google, it builds on the original MobileNet with enhancements like depth wise separable convolutions and an innovative inverted residual structure with linear bottlenecks. These features significantly reduce the computational burden while preserving accuracy. The architecture consists of convolutional layers followed by these separable convolutions, leading to a final fully connected layer that performs classification, as illustrated in the provided image. This design allows MobileNetV2 to achieve a balance between speed and performance, making it ideal for resource-constrained environments. (Bouteille, 2022)

- **VGG19 Model:**

VGG19 was developed by the Visual Geometry Group (VGG) at the University of Oxford, led by Karen Simonyan and Andrew Zisserman, and was introduced in their 2014 paper titled "Very Deep Convolutional Networks for Large-Scale Image Recognition."

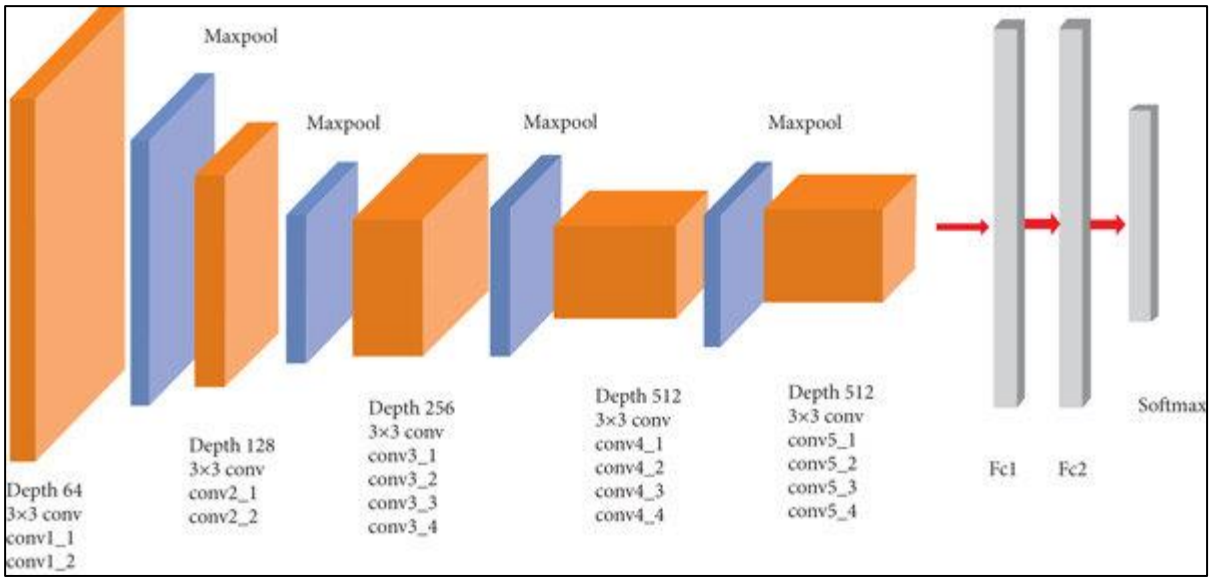


Figure 8. VGG19 Architecture
Source- *(Ullah, 2022)*

VGG19 is a deep convolutional neural network characterized by its simplicity and depth. It consists of five blocks of convolutional layers, each followed by max-pooling. The convolutional layers use small 3x3 filters, and the number of filters increases with each block (64, 128, 256, 512, 512). Following the convolutional layers are three fully connected layers, the first two with 4096 units each, and a final softmax layer for classification.

This architecture, shown in the accompanying diagram, is known for its consistent structure and has been highly influential in image classification tasks, where it achieves high accuracy through its deep and straightforward design. (geeksforgeeks, 2024)

- **Resnet152V2 Model:**

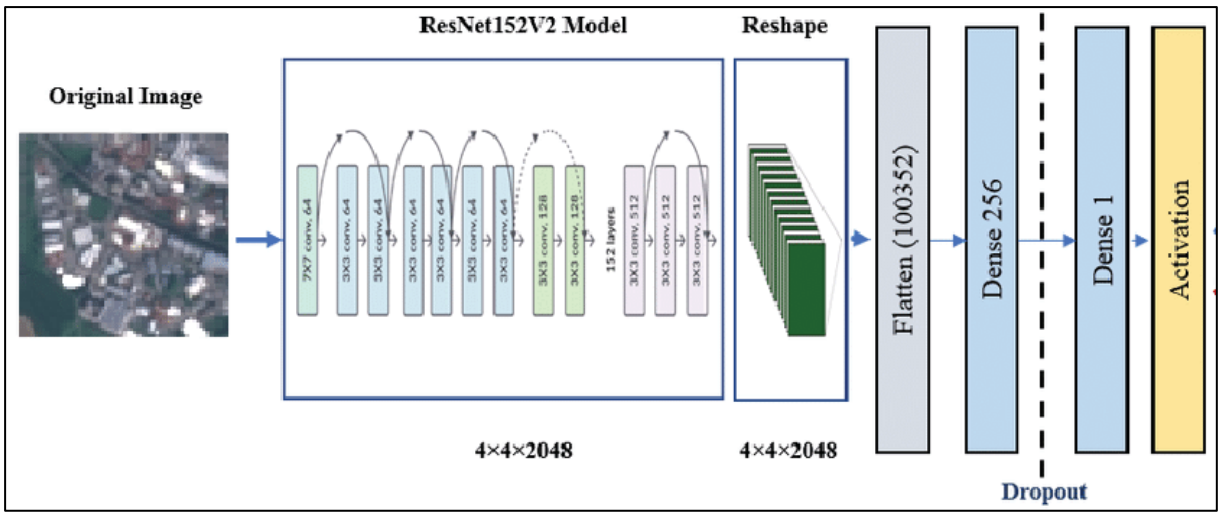


Figure 9. Resnet Model Architecture Source- *(Kittusamy, 2021)*

ResNet152V2 is a deep residual network, an advanced version of the original ResNet, designed to address the degradation problem in deep networks. Introduced by Kaiming He and his team at Microsoft Research, ResNet152V2 incorporates identity mappings within residual blocks, which helps in preserving gradients during backpropagation. The architecture is composed of a series of convolutional layers grouped into residual blocks, followed by batch normalization, ReLU activation, and skip connections. This structure allows for very deep networks to be trained more effectively without the vanishing gradient problem. The network concludes with global average pooling, followed by fully connected layers for classification. The diagram illustrates this architecture, showing the flow from the input image through convolutional and residual blocks to the final classification layers. (Kittusamy, 2021)

3.4 Segmentation and Survival Prediction

The U-Net architecture, introduced by Ronneberger, Fischer, and Brox in 2015, revolutionized biomedical image segmentation by enabling accurate results with minimal training data. Initially designed for 2D images, U-Net's success led to the development of the 3D U-Net, tailored for volumetric data like MRI and CT scans. The 3D U-Net processes 3D image cubes, capturing spatial relationships across all dimensions, making it highly effective for tasks requiring precise segmentation within a 3D space.

The architecture consists of symmetric downsampling and upsampling paths, enhancing its capability for detailed segmentation tasks. Here's a breakdown of the architecture:

The 3D U-Net architecture, designed for precise volumetric segmentation, consists of four main components: the contraction path, bottleneck, expansion path, and output layer. The contraction path reduces spatial dimensions while increasing feature channels through layers of 3D convolutions, ReLU activations, and max pooling, capturing increasingly abstract representations. The bottleneck layer, situated at the network's deepest point, serves as a bridge, condensing the input information. The expansion path reconstructs the spatial dimensions, using up-convolutions and skip connections to combine detailed low-level and high-level features from the contraction path. (Vinod, 2020)

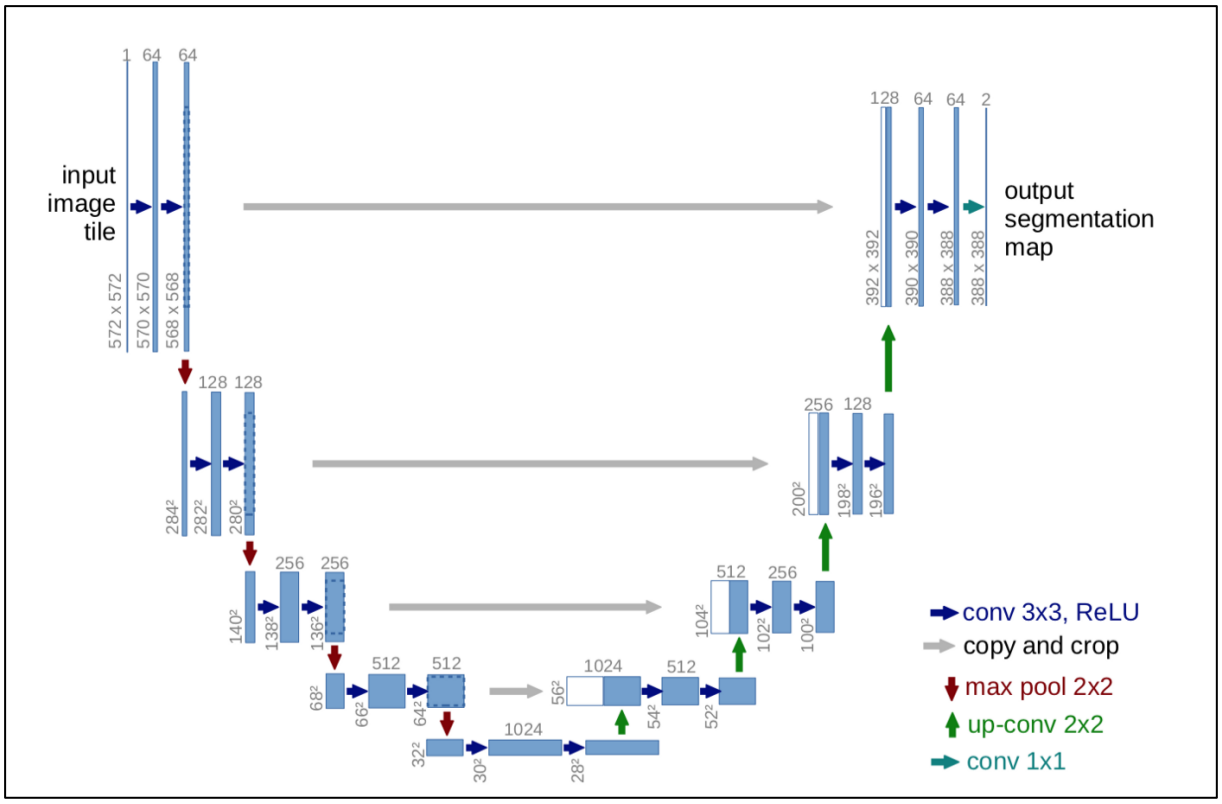


Figure 10. 3D UNET Architecture

Source- *(homola, 2018)*

This ensures accurate segmentation by leveraging both contextual information and fine details. The output layer, consisting of a 1x1x1 convolution followed by a softmax activation, produces a 3D segmentation map, assigning each voxel a probability for belonging to specific classes, such as tumor or non-tumor regions. This architecture is particularly effective in medical imaging, where accurate segmentation of complex structures is critical. (Vinod, 2020)

3D Attention U-Net: Extending the 3D U-Net, this architecture integrates attention gates in the skip connections between the contraction and expansion paths. These gates help the model focus on the most relevant features, suppressing less informative regions and enhancing the segmentation of crucial areas like the tumor core and edema. The attention mechanism, combined with the traditional U-Net structure, results in more accurate and focused segmentation, particularly beneficial in medical imaging where precision is critical. (Vinod, 2020)

3.5 Evaluation Metrics

Evaluation metrics are crucial in assessing the performance of machine learning models, as they provide a quantitative basis to determine how well a model is performing. Each metric serves a specific purpose and provides unique insights into model behaviours.

- **Intersection over Union (IoU) Score**

The Intersection over Union (IoU) score, also known as the Jaccard index, is a common metric used for evaluating the performance of segmentation models. It measures the overlap between the predicted segmentation and the ground truth, divided by the union of the predicted and ground truth regions. (Rosebrock, 2016)

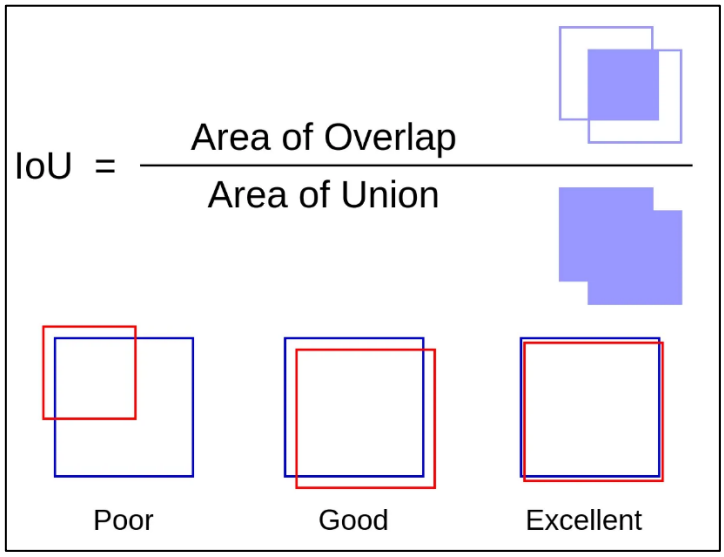


Figure 11. IoU Score Formula
Source- *(Rosebrock, 2016)*

- **Precision, Recall, and F1-Score**

Precision, recall, and F1-score are key metrics used to evaluate classification models. Precision measures the proportion of correctly predicted positive observations among all predicted positives, indicating the model's accuracy in identifying a specific class. Recall, or sensitivity, assesses the proportion of actual positives correctly identified by the model, reflecting its ability to capture all relevant instances. The F1-score, the harmonic mean of precision and recall, balances these two metrics, particularly in cases where the data is imbalanced. (Japkowicz, 2015)

$$Precision = \frac{True\ Positives\ (TP)}{True\ Positives\ (TP) + False\ Positives\ (FP)} \ldots\ldots\ldots \text{(Japkowicz, 2015)}$$

$$Recall = \frac{True\ Positives\ (TP)}{True\ Positives\ (TP) + False\ Negatives\ (FN)} \ldots.$$ (Japkowicz, 2015)

$$F1\ Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \ldots.$$ (Japkowicz, 2015)

- **Accuracy**

Accuracy is the most intuitive metric, representing the ratio of correctly predicted instances to the total instances in the dataset. While it is easy to understand, accuracy can be misleading when dealing with imbalanced datasets, as it does not differentiate between the types of errors made (false positives and false negatives). For this reason, while accuracy was reported in this study, more emphasis was placed on other metrics like precision, recall, and F1-score for the classification models. (Japkowicz, 2015)

$$Accuracy = \frac{True\ Positives\ (TP) + True\ Negatives\ (TN)}{Total\ Number\ of\ Samples} \ldots.$$ (Japkowicz, 2015)

- **Confusion Matrix**

A confusion matrix provides a more comprehensive evaluation by showing the actual versus predicted classifications in a matrix format. It breaks down the counts of true positives, true negatives, false positives, and false negatives, giving a complete picture of model performance for each class. This allows for the identification of specific types of errors made by the model, which can be crucial for improving model robustness. (Japkowicz, 2015)

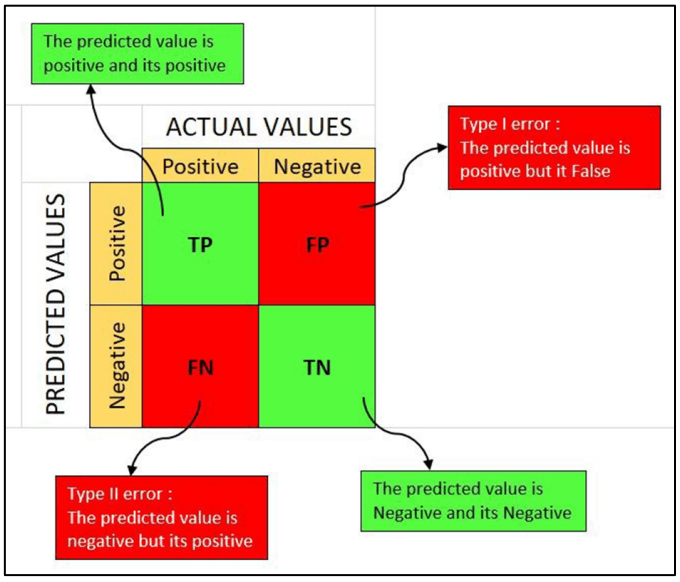


Figure 12. Confusion Matrix

Source *- (Naveen, 2022)*

CHAPTER 4

4. SYSTEM REQUIREMENTS

4.1 Datasets

Datasets play a pivotal role in the success of this project, serving as the foundation for both the brain tumor detection and survival prediction models. The quality, diversity, and comprehensiveness of the datasets directly influence the model's ability to learn and generalize effectively.

Due to the lack of a single comprehensive dataset that integrates both detection and survival prediction aspects, two separate datasets were utilized. This separation introduces challenges in terms of model integration and performance evaluation but allows for a focused approach to address the complexities of both tasks comprehensively. The integration of these datasets into the project is critical for achieving the dual objectives of accurate brain tumor detection and reliable segmentation and survival prediction, ultimately contributing to advancements in neuro-oncology research and clinical practice.

4.1.1 Brain Tumor Detection and classification dataset

The first dataset, obtained from Kaggle[1], comprises 7,023 human brain MRI images classified into four distinct categories: glioma, meningioma, no tumor (indicating the absence of tumors), and pituitary tumors. This dataset provided a rich source of labeled images for training and evaluating our brain tumor detection model. The diversity of the data in terms of tumor types is crucial for developing a robust classification system capable of accurately identifying various forms of brain tumors.

4.1.2 Segmentation and Survival Prediction Dataset

The second dataset, also sourced from Kaggle[2], is dedicated to the Brain Tumor Segmentation and survival prediction, includes multimodal MRI scans stored as NIfTI files (.nii.gz). These scans encompass native (T1), post-contrast T1-weighted (T1Gd), T2-weighted (T2), and T2 Fluid Attenuated Inversion Recovery (T2-FLAIR) volumes. Annotations for this dataset

[1] https://www.kaggle.com/datasets/masoudnickparvar/brain-tumor-mri-dataset
[2] https://www.kaggle.com/datasets/awsaf49/brats20-dataset-training-validation

include GD-enhancing tumor (ET), peritumoral edema (ED), and necrotic and non-enhancing tumor core (NCR/NET), facilitating the development of our survival prediction model. The rich annotation details and multimodal imaging data allow for comprehensive feature extraction necessary for accurate survival analysis.

4.2 Computational Environment

4.2.1 Spyder IDE

The Spyder IDE was chosen for implementing the machine learning models due to its ability to handle the computational demands that surpass the capabilities of cloud platforms like Google Colab. Spyder provides a stable environment for running resource-intensive models, avoiding the performance issues often encountered with Colab. It offers a comprehensive development environment with a powerful code editor, variable explorer, and interactive console, ideal for debugging and refining models. Additionally, by leveraging local hardware, Spyder optimizes GPU and CPU usage, resulting in faster model training and evaluation.

4.3 Computational Challenges

The study encountered significant computational challenges due to the large datasets and the intensive nature of the models utilized. The brain tumor classification dataset, though relatively small at 150 MB, required efficient processing to manage thousands of images. In contrast, the segmentation and survival prediction dataset, totalling 47 GB of multimodal MRI data, imposed considerable demands on storage, memory, and processing power.

The project was executed on a system with an AMD Ryzen 5 5500U processor featuring 12 cores at ~2.1 GHz, 16 GB of RAM, and Windows 11. Despite these specifications, the high computational requirements of the 3D U-Net and 3D Attention U-Net models resulted in extended training times—up to 48 hours for the 3D U-Net and 96 hours for the 3D Attention U-Net to complete 100 epochs. These durations underscore the need for more powerful computational resources, including high-end GPUs, increased RAM, and faster processors, to efficiently train these advanced models.

CHAPTER 5

5. METHODOLOGY

5.1 Brain Tumor Classification

5.1.1 Data Exploration

Data exploration is a crucial initial step in any machine learning project, as it provides a thorough understanding of the dataset, uncovers patterns, and prepares the data for subsequent modelling. In this study, the data exploration phase was carefully executed to gain detailed insights into the dataset. This involved tasks such as data loading, label generation, sample image visualization, and analysis of label distribution across classes.

Essential libraries like os, pandas, tensorflow, matplotlib, and seaborn were utilized for data manipulation, image processing, and visualization throughout this phase.

- **Dataset Overview**

The dataset comprised a total of 7023 images, categorized into four distinct classes: glioma, meningioma, pituitary, and no tumor. These images have already been divided into training and testing sets to facilitate the development and evaluation of machine learning models.

The training set distribution contained a total of 5734 images, distributed as follows:

- Glioma: 1321 images
- Meningioma: 1339 images
- Pituitary: 1457 images
- No Tumor: 1595 images

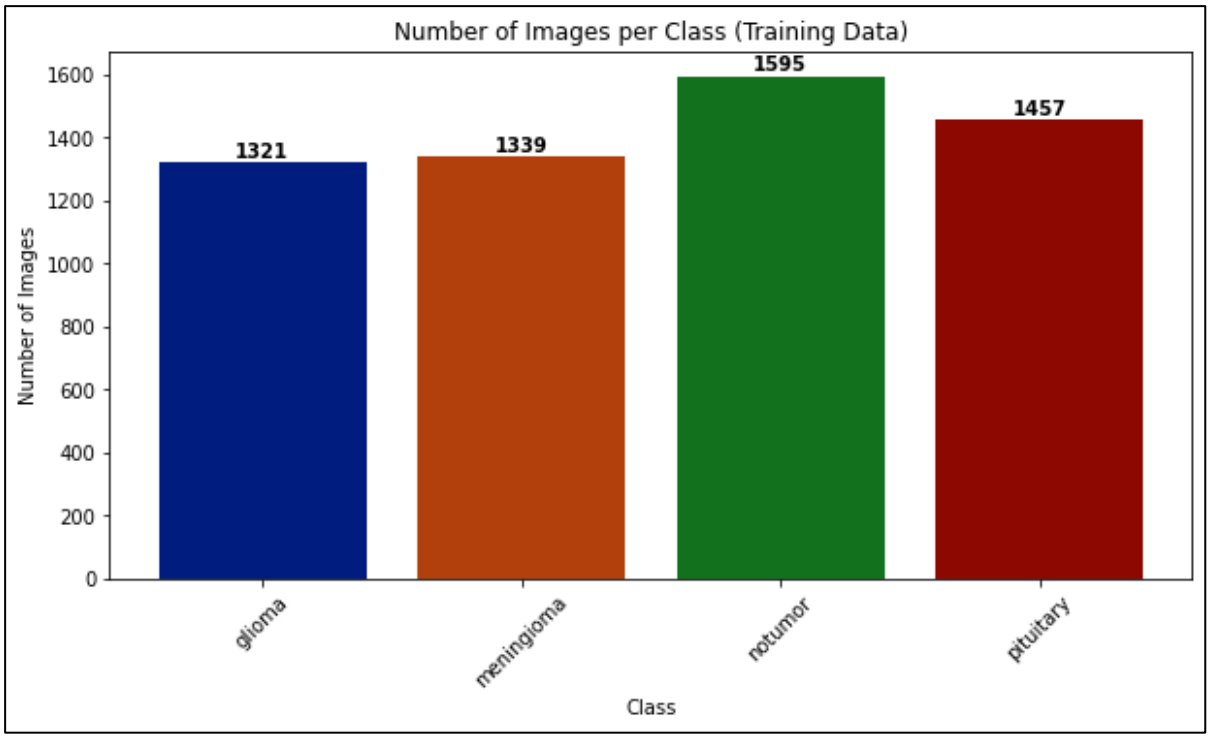


Figure 13. Distribution Of Train Set

The testing set distribution contained a total of 1311 images, distributed as follows:

- Glioma: 300 images
- Meningioma: 306 images
- Pituitary: 300 images
- No Tumor: 405 images.

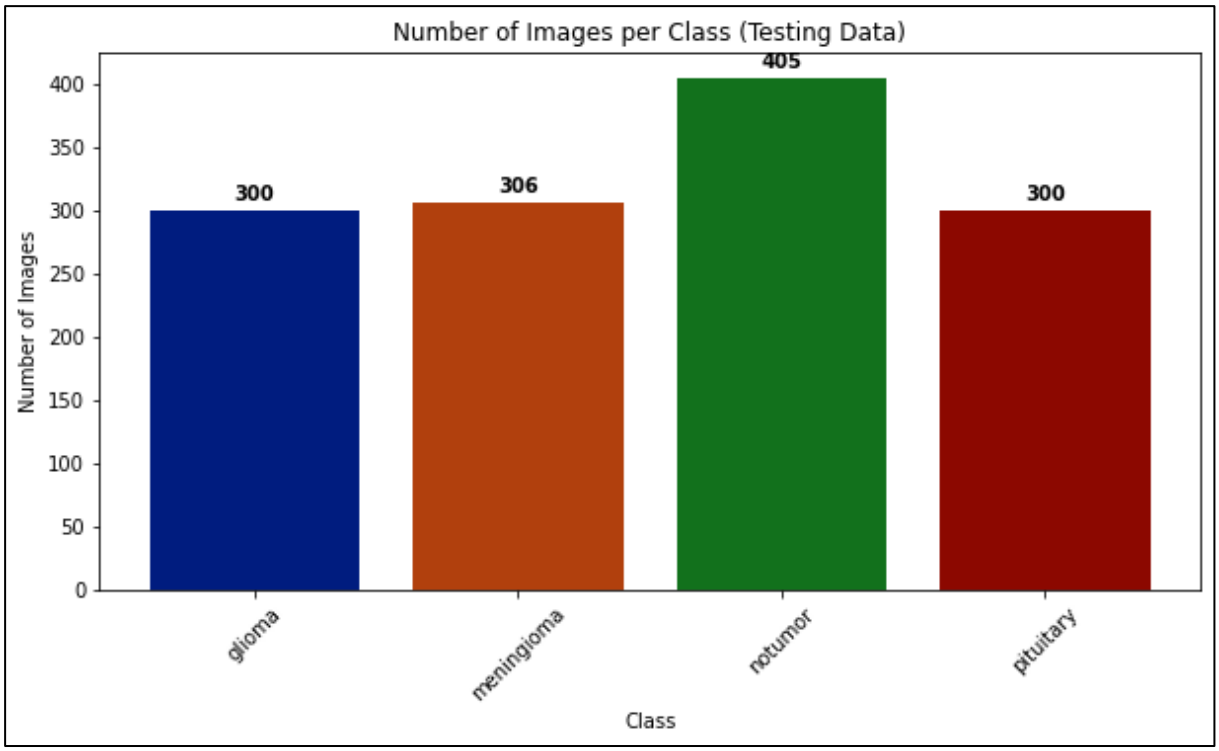


Figure 14. Distribution Of Test Set

The bar plots in **Figure 13** & **Figure 14** clearly indicated that the dataset was relatively balanced across the four classes, with a slightly higher representation of the "No Tumor" class. This balance was advantageous as it contributed to training models that were less biased towards any particular class, thereby ensuring more reliable predictions.

To gain deeper insights into the dataset, a selection of sample images from each class was visualized. **Figure 15** illustrated randomly selected samples from the dataset, with each row representing a different tumor type or the "No Tumor" category.

These visualizations provided a clear representation of the variations and complexities within the dataset. By examining these images, it was possible to observe the differences in tumor appearance, which underscored the challenges involved in accurate classification.

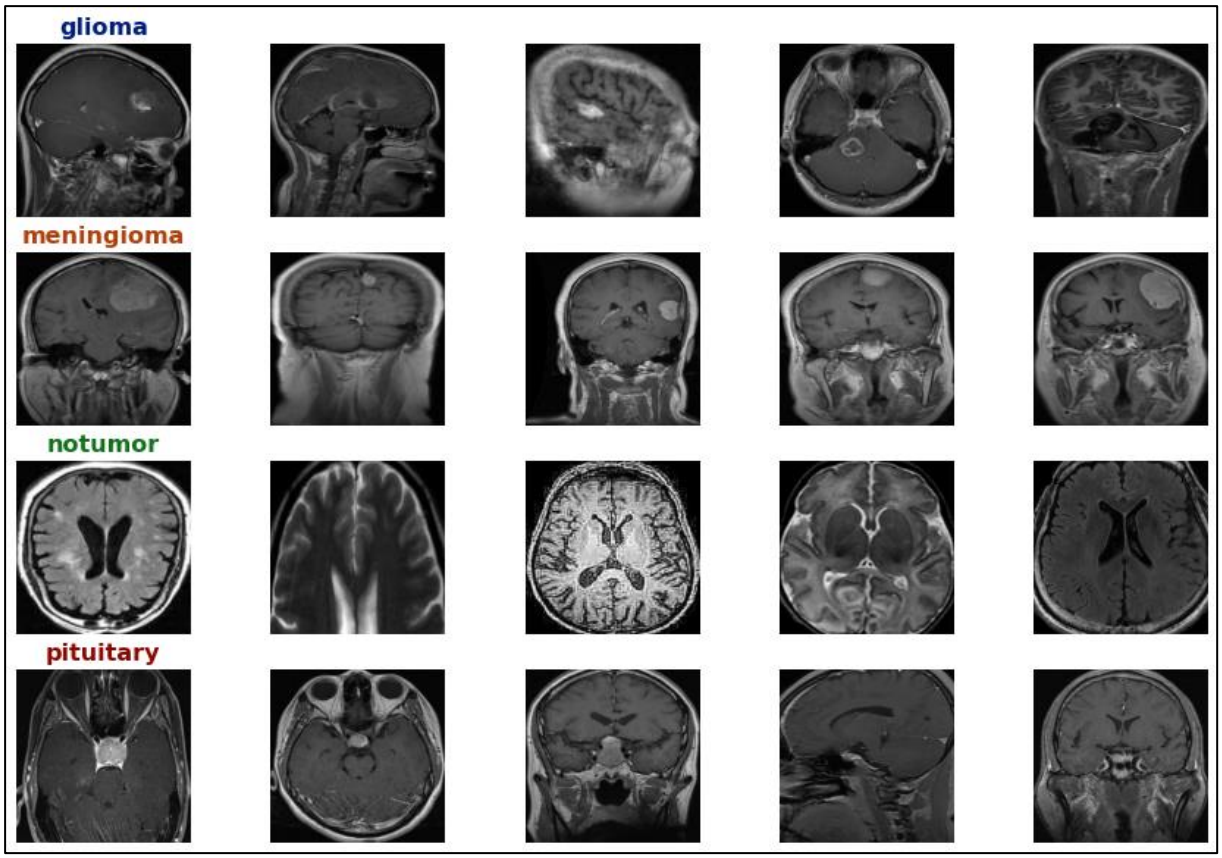


Figure 15. Sample Images from the Dataset

5.1.2 Data Preprocessing

Data preprocessing was a crucial step in preparing the dataset for modeling, as it enhanced the quality of the data and ensured that it was in a suitable format for machine learning algorithms. The preprocessing phase in this study involved several key operations, including image cropping, resizing, and augmentation.

- **Image Cropping and Resizing**

Given the variability in the size and content of the images, it was essential to standardize the dataset by cropping and resizing the images to a consistent size. This process involved the following steps:

- Gray-Scale Conversion: Images were converted to grayscale to simplify processing and emphasize structural details rather than color.
- Blurring and Thresholding: Applied Gaussian blur to reduce noise, followed by thresholding to separate the brain region from the background.

- Contour Detection: After thresholding, contours were detected in the binary image, isolating the brain by detecting the largest contour.
- Extreme Point Extraction: The extreme points (leftmost, rightmost, topmost, and bottommost) of the largest contour were determined. These points were used to define a rectangular region that tightly enclosed the brain, effectively cropping out irrelevant parts of the image.
- Cropping: Using the extreme points, the images were cropped to focus solely on the brain region, removing irrelevant background.
- Resizing: The cropped images were then resized to a uniform size of 256x256 pixels, ensuring consistency for neural network input.
- Saving the Preprocessed Images: Organized and saved processed images in directories by class, streamlining access for model training.

- **Data Augmentation**

To improve the generalization ability of the model and prevent overfitting, data augmentation techniques were applied to the training dataset. Data augmentation artificially expanded the dataset by generating new training examples through various transformations. The following augmentation techniques were employed:

- Rescaling**:** All pixel values in the images were rescaled by a factor of 1/255. This normalization step ensured that the pixel values fell within the range [0, 1], which was necessary for faster convergence during model training.
- Shear Transformation**:** A shear range of 0.2 was applied to introduce affine transformations, where the image was slanted along the x or y-axis. This transformation helped the model become invariant to small distortions in the images.
- Zooming: A zoom range of 0.2 was utilized to randomly zoom in on the images. Zooming aided in teaching the model to recognize objects (tumors) at various scales.
- Horizontal and Vertical Flipping: The images were randomly flipped both horizontally and vertically. Flipping was particularly useful for medical images as it increased the diversity of the training data without altering the underlying content.
- Fill Mode**:** When transformations resulted in pixels being shifted outside the image boundaries, the empty areas were filled using the nearest pixel values. This ensured that the augmented images remained realistic.

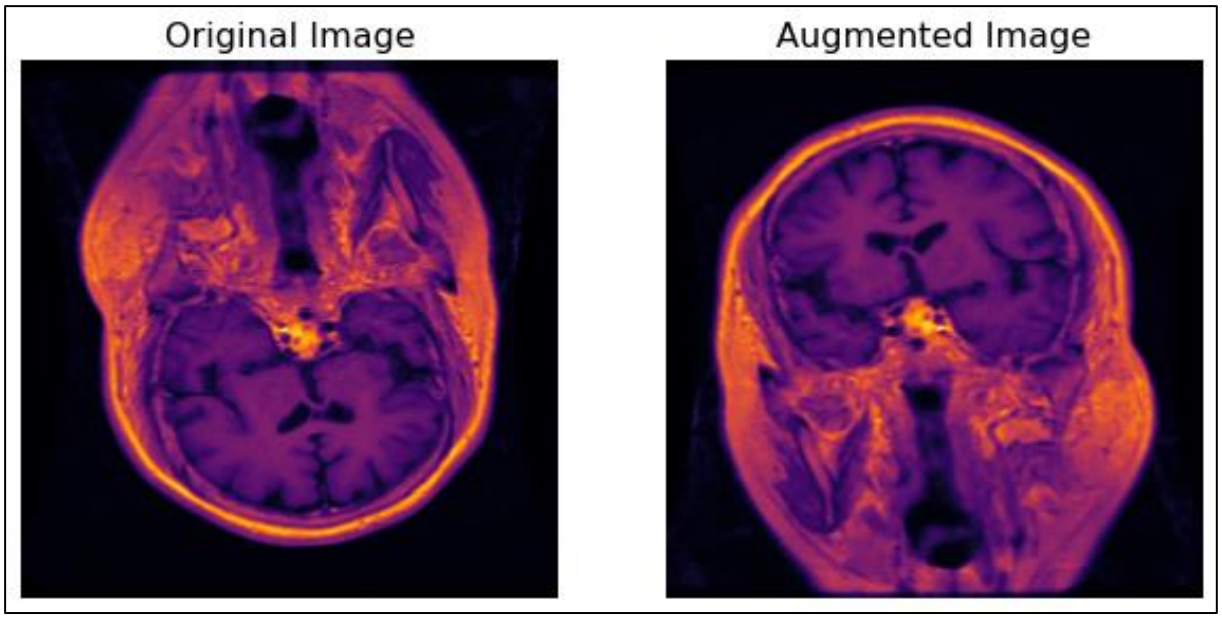


Figure 16. Comparison Between Original and Augmented Images

From the **Figure 16**, it was evident that the augmentation layer generated images that were slightly different from the original images. This was intentional, as the goal was to produce augmented images that were similar to the original dataset while preserving the key features of the images. Data augmentation was critical as it helped improve the robustness and generalization capability of the model by exposing it to a wider range of scenarios during training.

- **Dataset Preparation**

Following data augmentation, the dataset was prepared for training, validation, and testing:

Training Data: The augmented images were loaded in batches from the training directory, with a target size of 256x256 pixels and a batch size of 32. The class_mode was set to categorical, meaning that the output labels were one-hot encoded, which was necessary for multi-class classification tasks like this one.

Validation and Testing Data: Both the validation and testing datasets were processed similarly. Rescaling was applied to normalize the pixel values in these datasets. The images were loaded in batches from their respective directories, with a target size of 256x256 pixels and a batch size of 32. The class_mode for both the validation and testing datasets was set to categorical, ensuring consistency in the evaluation process.

The use of an ImageDataGenerator for training, validation, and testing datasets ensured that the images were processed on the fly during model training and evaluation, reducing the memory load and allowing for efficient data handling. By applying these preprocessing steps, the model was better equipped to learn from the data and perform well on unseen examples.

5.1.3 Model Building

The approach for building each model was consistent, leveraging the power of transfer learning. Pre-trained models such as InceptionV3, MobileNetV2, ResNet152V2, and VGG19, each with weights learned from the ImageNet dataset, were utilized to form the backbone of the classification system. The initial layers of these models were frozen to retain the powerful features learned from extensive prior training.

Custom Classification Head: For each model, a custom classification head was appended. This head typically included a global average pooling layer, a dense layer with 256 units and ReLU activation, followed by a softmax layer to output the probability of each of the four classes. This common structure across models ensured consistency in training and evaluation.

Training Process: Each model was compiled using the Adam optimizer and categorical cross-entropy loss function. Training was conducted across multiple epochs, during which accuracy and loss metrics were monitored closely. Adjustments to the learning rate were made dynamically to optimize performance.

Model-Specific Details

InceptionV3: This model, with its depth and inception modules, was trained for 20 epochs. It was specifically chosen for its capability to handle intricate image features.

MobileNetV2: Recognized for its efficiency, particularly in mobile and edge applications, MobileNetV2 was trained over 10 epochs. Its architecture is lightweight yet effective, making it a strong contender for resource-constrained environments.

ResNet152V2: Known for its deep residual connections, ResNet152V2 was trained for 15 epochs. Its architecture excels at maintaining accuracy even as the depth of the network increases.

VGG19: With a focus on simplicity and depth, VGG19 was also trained over 10 epochs. Its straightforward architecture makes it a reliable choice for various image classification tasks.

Evaluation and Outcomes

Post-training, each model's performance was evaluated using accuracy, loss plots, confusion matrices, and classification reports. The use of transfer learning not only accelerated the training process but also enabled high accuracy in classifying brain tumors across all models. The results from each model were saved for future inference and potential ensemble methods.

5.1.4 Ensemble Techniques

To further enhance the robustness and accuracy of the brain tumor classification system, ensemble learning techniques were implemented. Ensemble learning combines the strengths of multiple models, thereby mitigating the weaknesses inherent in individual models. In this study, three ensemble methods were applied—simple averaging, weighted averaging, and geometric mean. These techniques aggregated the predictions from several pre-trained models, including InceptionV3, MobileNetV2, ResNet152V2, and VGG19, to generate a more robust and accurate final classification.

- **Simple Average Ensemble**:

The simple average ensemble technique combines the predicted probabilities from all the individual models (InceptionV3, MobileNetV2, ResNet152V2, and VGG19) by computing the arithmetic mean. This method assumes that all models contribute equally to the final prediction. The averaged probabilities are then used to make the final class prediction by selecting the class with the highest average probability. This approach generally enhances performance by reducing the variance associated with any single model's predictions.

- **Weighted Average Ensemble**:

In the weighted average ensemble, different weights were assigned to each model's predictions based on their individual performance metrics. This technique allows for emphasizing more reliable models while down-weighting those that may not perform as well. By multiplying the predicted probabilities by their respective weights and summing them, a weighted average

probability is obtained. The final class prediction is made based on these weighted probabilities, providing a more balanced and informed decision-making process that capitalizes on the strengths of the better-performing models. To further explore the impact of weighting, equal weights were assigned to all models. Interestingly, this adjustment did not lead to significant changes in the overall performance metrics, suggesting that the models contributed relatively equally to the ensemble's effectiveness. This finding underscores that, in this case, careful tuning of weights might not be as crucial as initially anticipated, although it can still lead to marginal gains.

- **Geometric Mean Ensemble**:

The geometric mean ensemble method computes the geometric mean of the predicted probabilities from each model. This method is particularly effective when models have varying levels of confidence in their predictions, as it is less sensitive to extreme values than the arithmetic mean. The geometric mean ensemble provides a robust combination of the models' outputs, often leading to improved performance in scenarios where predictions need to be stabilized across diverse models.

5.2 Brain Tumor Segmentation and Survival Prediction

5.2.1 Data Exploration

In the data exploration phase of the brain tumor segmentation and survival prediction project, a comprehensive approach was adopted to understand its structure and content of the MRI data. The dataset comprised multimodal scans stored in NIfTI format (.nii.gz), a standard format in medical imaging, particularly for brain MRI scans. The different MRI modalities included T1, T1c, T2, and FLAIR images, each offering unique insights and perspectives on brain tissue and tumor characteristics.

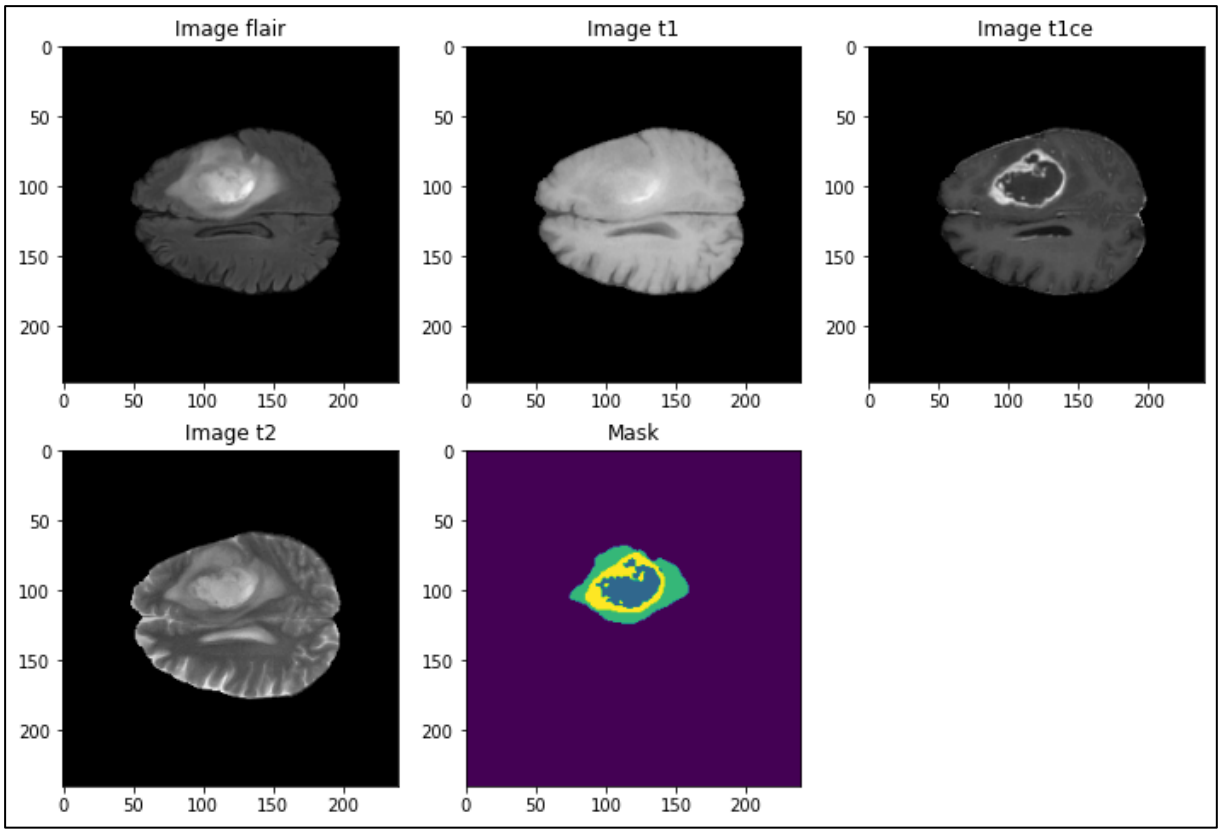


Figure 17. Sample from Brain Tumor segmentation Dataset

The **Figure 17**, shows how each modality provides a different perspective of the brain tissue, with the segmentation mask clearly delineating the tumor regions.

Each MRI modality serves a distinct purpose:

- **T1**: T1-weighted images, often acquired using sagittal or axial 2D techniques, with a slice thickness ranging from 1 to 6 mm, provide detailed anatomical information. (BraTS2020 Dataset (Training + Validation), 2020)
- **T1c**: T1-weighted, contrast-enhanced (Gadolinium) images are typically acquired using 3D techniques, offering a 1 mm isotropic voxel size for most patients. These images enhance the visibility of the tumor by highlighting areas where the contrast agent accumulates.
- **T2**: T2-weighted images, generally acquired using axial 2D techniques with a slice thickness of 2 to 6 mm, are particularly useful for detecting edema and changes in water content within the brain tissue.

- **FLAIR**: T2-weighted FLAIR images, which can be acquired in axial, coronal, or sagittal planes with a slice thickness of 2 to 6 mm, suppress the cerebrospinal fluid signal, making it easier to detect lesions near the fluid spaces.

Each image was loaded using the NiBabel library, which is well-suited for handling NIfTI files. A sample image was selected from the dataset for visualization, which involved loading and scaling the image data. The T1, T1ce, T2, and FLAIR images, along with their corresponding segmentation masks, were visualized to inspect the quality and consistency of the data. The segmentation masks, which originally had pixel labels of 0, 1, 2, and 4, were adjusted by reassigning the label 4 to 3 to maintain consistency across the dataset. This re-encoding was essential for the segmentation task, where the classes were to be treated as distinct entities. (BraTS2020 Dataset (Training + Validation), 2020) The segmentation masks classified the brain regions into four distinct categories:

- **Class 0: 'NOT tumor'** – healthy brain tissue,
- **Class 1: 'CORE'** – necrotic or non-enhancing tumor core,
- **Class 2: 'EDEMA'** – the area of swelling surrounding the tumor,
- **Class 3: 'ENHANCING'** – the active, contrast-enhancing tumor region.

5.2.2 Data Preprocessing

Following the exploration, the data preparation phase focused on preprocessing the images and masks for model training. This included scaling the images using the MinMaxScaler to normalize the pixel intensity values, ensuring that all images were within a uniform range. The different modalities were combined into a single multi-channel image, effectively consolidating the different perspectives of the MRI scans into one dataset. The dataset comprised three MRI modalities: FLAIR, T1ce, and T2, which were used as input channels, also seen in the below in **Figure 18**. The T1 modality was excluded from the inputs because the T1ce images provided the necessary contrast information, rendering T1 redundant, which can also be seen in **Figure 17**.

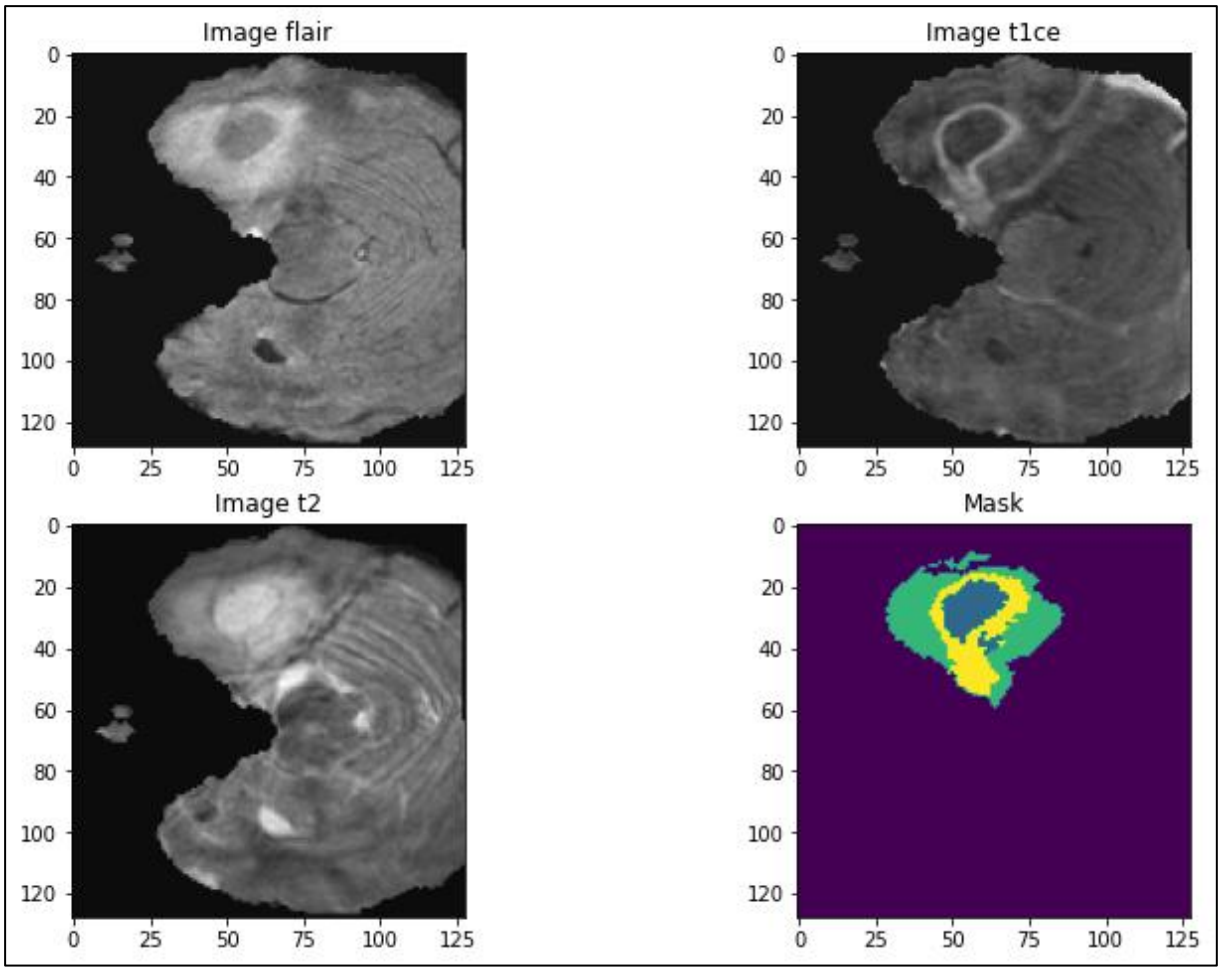


Figure 18. Input Image for the models

To ensure that the images were compatible with the input requirements of the segmentation models, the combined images were cropped to dimensions divisible by 64, resulting in 128x128x128 pixel volumes. This allowed for the extraction of sub-volumes or patches during training. The corresponding segmentation masks were similarly processed to match the dimensions of the images. The cropped and combined images, along with their masks, were then saved in both TIFF and Numpy formats for ease of use in the subsequent modeling phase.

The dataset was further segregated into training and validation sets to facilitate model training and evaluation. This segregation was crucial for ensuring that the model's performance could be accurately assessed on unseen data. The training data was split using an 80:20 ratio, with 80% of the data allocated for training and 20% for validation. The split-folders library was utilized to automate this process, ensuring that the images and masks were organized into the appropriate directories.

To maintain consistency, the filenames of the images and masks were carefully checked to ensure that each image had a corresponding mask. This verification step was necessary to prevent mismatches during model training, which could lead to errors or reduced model performance. The data preparation process concluded with the organization of the training and validation datasets into separate directories, ready for the subsequent phases of model training and evaluation.

5.2.3 Brain segmentation

In the development of the brain tumor segmentation model, a comprehensive approach was adopted, focusing on leveraging 3D U-Net architecture. This method was chosen due to its effectiveness in handling volumetric data, such as MRI scans. The process involved several key steps, which are outlined below.

Data Preparation and Generator Creation

The model building process began with the preparation of the dataset, where MRI images and corresponding segmentation masks were loaded and pre-processed. To manage the large volume of data, an image loader generator was created. This generator was responsible for dynamically loading the images and masks in batches during the training process. It ensured that the images and masks were correctly paired and appropriately scaled, facilitating efficient memory usage and allowing for the model to be trained on large datasets without encountering memory constraints.

3D U-NET Model Architecture

The 3D U-Net architecture, as implemented, adhered to the traditional U-Net structure with tailored modifications for volumetric data processing. The model began with an input layer designed to accept 128x128x128 voxel data.

The **encoder path** consisted of five convolutional blocks, each progressively more complex in terms of the number of filters. The blocks began with 16 filters and doubled up to 256 filters as the network deepened. Each block utilized 3x3x3 convolutional layers with ReLU activations, followed by max-pooling layers that downsampled the spatial dimensions. Dropout layers were

incorporated to prevent overfitting during training. The encoder path effectively captured increasingly abstract features as the data moved deeper into the network.

At the deepest part of the network**, the bottleneck layer** consisted of two convolutional layers with 256 filters each. This bottleneck layer captured the most abstract and high-level features from the input data before passing the information to the decoder path.

The **decoder path** mirrored the encoder path but utilized transposed convolutions to upsample the spatial dimensions. Skip connections between corresponding encoder and decoder blocks were essential for preserving spatial details that were lost during downsampling. These connections allowed the network to combine high-level features from the bottleneck with preserved spatial information from the encoder, which was crucial for accurate segmentation.

Activation Function (ReLU): The 3D U-Net employed the ReLU (Rectified Linear Unit) activation function in its convolutional layers. ReLU was chosen for its simplicity and effectiveness in deep learning models. It introduces non-linearity by setting all negative values to zero, which helps the network learn complex patterns. ReLU accelerates convergence during training and mitigates the vanishing gradient problem, making it well-suited for deep networks like the 3D U-Net.

The model concluded with a softmax-activated 1x1x1 convolution layer, which produced a probability map for each voxel, allowing the network to classify each voxel into distinct categories.

The **He uniform initializer** was employed for weight initialization, ensuring a faster and more stable convergence during the training process. This architecture was carefully designed to maximize segmentation accuracy while efficiently handling the challenges of volumetric data.

Model Compilation and Training

The 3D U-Net was compiled using a combination of Dice Loss and Categorical Focal Loss. These loss functions were specifically selected to address class imbalance and to focus learning on difficult-to-classify examples.

- **Dice Loss:** This loss function optimized the overlap between predicted and actual segments, making it highly effective for medical image segmentation tasks where class imbalance is common.
- **Categorical Focal Loss:** This loss function adjusted the learning process to emphasize difficult cases, ensuring that the model did not overly favor the majority class.

The Adam optimizer, with a learning rate of 0.0001, was employed for training. The metrics for model evaluation included accuracy and the Intersection over Union (IoU) score, which are crucial for measuring segmentation performance. The model was trained over 100 epochs with a batch size of 2, utilizing the image loader generator for both the training and validation datasets. After training, the model was saved for future use, and the training history, including loss and accuracy metrics, was plotted to assess the model's performance over time.

3D Attention U-Net Architecture

The 3D Attention U-Net was developed as an enhancement of the traditional 3D U-Net, incorporating advanced attention mechanisms to improve segmentation accuracy. Attention mechanisms allow the model to focus on the most relevant regions in the MRI scans, ensuring that the segmentation process is both precise and accurate.

Data Preparation and Augmentation

The data preparation for the 3D Attention U-Net followed the same process as the 3D U-Net, using the same MRI modalities and cropping the images to 128x128x128. In addition to this, the image loader generator was enhanced with data augmentation techniques, such as random rotation and flipping. These augmentations improved the model's robustness by making it more adaptable to variations in the data, thereby enhancing its generalization capabilities.

Model Architecture

The **encoder path** of the 3D Attention U-Net was similar to that of the 3D U-Net, with convolutional blocks that increased in complexity from 16 to 256 filters. However, instead of ReLU activations, the 3D Attention U-Net employed Leaky ReLU activations. Leaky ReLU was chosen over ReLU because it allows a small gradient to pass through even when the input is negative, preventing the "dying ReLU" problem where neurons could become inactive. This

was particularly important in the attention mechanism, where retaining subtle features could be critical for accurate segmentation.

The **bottleneck** in the 3D Attention U-Net, located at the deepest layer of the encoder, consisted of two convolutional layers with 256 filters each. This section of the network captured the most abstract features from the input data, setting the stage for the decoding process. The bottleneck was essential for summarizing the learned features before they were progressively upsampled in the decoder path.

- **Gating Signal:** Gating signals were a key component of the attention mechanism in the 3D Attention U-Net. These signals were generated using 1x1x1 convolutions followed by batch normalization and ReLU activation. Derived from the deeper layers of the network, the gating signals provided contextual information that guided the attention mechanism in focusing on the most relevant features. By modulating the attention blocks with these signals, the model was able to selectively emphasize important regions while downplaying irrelevant ones. ReLU was used in the gating signal because it effectively filters out irrelevant features by zeroing out negative values, ensuring that only the most important information is passed to the attention mechanism.
- **Attention Block:** The attention blocks were another critical innovation in the 3D Attention U-Net. Each attention block received inputs from both the encoder (providing spatial information) and the gating signals (providing contextual information). These blocks computed attention maps that highlighted the relevant regions in the feature maps. The attention maps were then used to modulate the encoder's feature maps, enhancing the important regions and suppressing the irrelevant ones. This modulation ensured that the features passed through the skip connections were focused and relevant, leading to more accurate segmentation results.
- **Dual Attention Mechanism:** The core innovation in the 3D Attention U-Net was the introduction of a dual attention mechanism at multiple stages of the decoder path. This mechanism included both channel attention and spatial attention.
    - Channel Attention: This mechanism focused on selecting the most important feature maps among all the feature maps. By applying channel attention, the network effectively prioritized certain features over others, ensuring that the most relevant features for segmentation were given more weight.

- Spatial Attention: Spatial attention highlighted the most critical regions within the feature maps, directing the network's focus to the spatial locations that mattered most. This was particularly important for identifying tumor regions within the MRI scans.

These attention mechanisms were implemented using attention blocks that modulated the features passed through the skip connections between the encoder and decoder paths. The attention blocks ensured that only the most relevant features were preserved and passed to the decoding stages, enhancing the network's ability to accurately segment the brain tumors.

The **decoder path** in the 3D Attention U-Net was similar to that of the 3D U-Net but with the crucial addition of attention-modulated skip connections. These connections allowed the model to focus on the most relevant features during the upsampling process, improving the accuracy of the segmentation.

The **output layer** was a 1x1x1 convolutional layer with a softmax activation function, producing a voxel-wise probability map. This allowed the network to classify each voxel into the appropriate category, ensuring accurate and detailed segmentation.

The 3D Attention U-Net utilized custom loss functions tailored to the challenges of medical image segmentation.

- **Dice Coefficient Loss** focused on optimizing the overlap between predicted and actual segmentation masks, effectively addressing class imbalance and ensuring accurate segmentation of even small tumor regions.
- **Tversky Loss** extended the Dice loss by allowing asymmetric penalties on false positives and false negatives, making it suitable for scenarios where certain types of errors need to be minimized.
- **Focal Tversky Loss** further refined the Tversky loss by emphasizing difficult-to-segment regions, enhancing the model's performance in challenging cases like small or diffuse tumors.

Model Compilation and Training

The 3D Attention U-Net was compiled using the Focal Tversky Loss, which is particularly effective in handling class imbalance and challenging segmentation scenarios. The Adam optimizer, with a learning rate of 0.0001, was used to train the model.

The model was trained for 100 epochs with a batch size of 2, similar to the 3D U-Net. The image loader generator, enhanced with data augmentation techniques, was used to improve the model's robustness. The training process involved monitoring the validation loss to save the best model, ensuring that the model did not overfit to the training data. The training history, including loss, accuracy, and IoU metrics, was plotted to evaluate the model's performance over time.

5.2.4 Survival prediction

Data Exploration and Preparation

The survival prediction task focused on estimating patient survival duration based on brain tumor characteristics using data from the BraTS 2020 challenge. Survival days were categorized into three classes: Short (< 250 days), Medium (250-450 days), and Long (> 450 days). The primary challenge involved extracting features from segmented tumor regions in MRI scans that could correlate with survival outcomes.

The dataset was meticulously pre-processed to include only entries with corresponding MRI scans, segmentation masks, and T1-weighted images. This ensured the accuracy of volumetric features calculated for various tumor regions—such as necrotic core, edema, and enhancing tumor—which were normalized by total brain volume. These volumetric features, along with patient age and encoded ResectionStatus, were used as input features for the models, capturing both anatomical and clinical aspects of the tumors.

The boxplot of survival days **Figure 19** shows that most patients have survival times clustered around the "Short" (<250 days) and "Medium" (250-450 days) categories, with fewer in the "Long" (>450 days) category. There are several outliers indicating very long survival times (>1000 days), suggesting variability in patient outcomes based on underlying factors such as tumor characteristics, patient demographics, and treatments received.

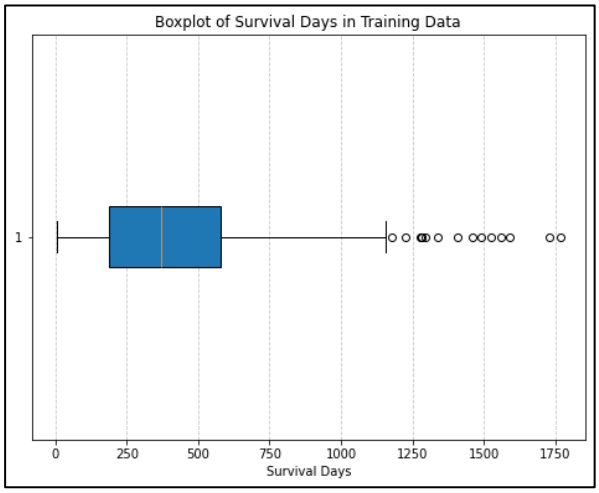


Figure 19. Box Plot for Target Variable (Survival Prediction)

The correlation matrix **Figure 20** shows that:

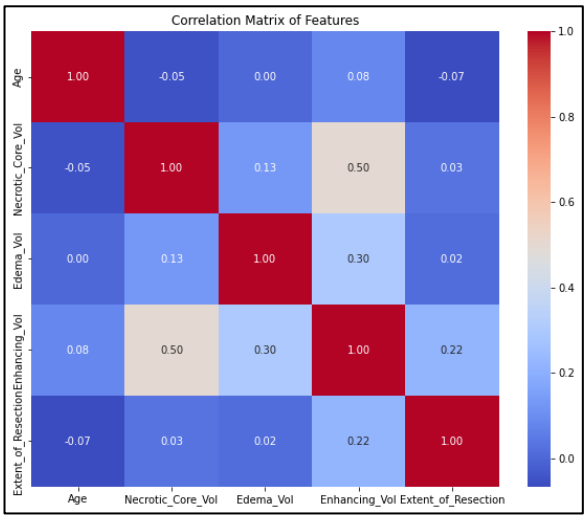


Figure 20. Correlation Matrix Survial Prediction Features

Age has a weak negative correlation with survival-related features, indicating age alone is not a strong predictor of survival. **Necrotic Core Volume** and **Enhancing Tumor Volume** are moderately correlated (0.50), suggesting larger volumes are associated with more aggressive tumors and shorter survival. **Edema Volume** has low correlations with other features, implying it is not a strong standalone predictor and might impact survival only when combined with other factors. **Extent of Resection** is weakly correlated with Enhancing Tumor Volume (0.22), hinting that larger tumors may require more extensive surgery, potentially improving survival.

The data was split into training and validation sets in an 80:20 ratio. The Synthetic Minority Over-Sampling Technique (SMOTE) was applied to the training set to address class imbalance, preventing model bias toward the majority class. Data standardization using StandardScaler ensured consistency in feature scales, which is crucial for models like SVMs that are sensitive to input magnitudes.

Hyperparameter optimization was performed using RandomizedSearchCV to efficiently find optimal parameters within large search spaces. Unlike Grid Search, RandomizedSearchCV sampled random combinations, enabling it to identify optimal settings with fewer iterations (Koehrsen, 2018). Cross-validation was incorporated to ensure robust model tuning and minimize overfitting. The use of SMOTE further enhanced the models by refining them to produce reliable performance metrics, improving generalization across varied scenarios.

Model Building

Three primary machine learning models were developed for the survival prediction task: Random Forest, Gradient Boosting, and Support Vector Machine (SVM). Each model underwent hyperparameter tuning to optimize their performance, leveraging **RandomizedSearchCV** to efficiently search across a range of hyperparameters and identify the best combination.

- Random Forest Classifier

Random Forest is an ensemble learning method that uses multiple decision trees to predict the output class. It reduces the risk of overfitting and generally results in better accuracy and stability.

Table 1. Hyperparameter Tuning for Random Forest

Hyperparameter	Values	Description
n_estimators	randint(50, 100)	Number of trees in the forest, affecting model complexity and generalization ability. (Koehrsen, 2018)
max_depth	randint(3, 10)	Maximum depth of each tree, controlling the model's capacity to capture complex relationships in the data. (Koehrsen, 2018)
min_samples_split	randint(5, 20)	Minimum number of samples required to split an internal node, influencing the model's robustness. (Koehrsen, 2018)

min_samples_leaf	randint(5, 20)	Minimum number of samples required to be at a leaf node, affecting the model's ability to generalize. (Koehrsen, 2018)

Best parameters for Random Forest Classifier --

max_depth: 6 , min_samples_leaf: 6 , min_samples_split: 14, n_estimators: 94.
Best cross-validation score for Random Forest: 0.49279907084785135

- Gradient Boosting Classifier

Gradient Boosting builds an additive model in a forward stage-wise manner; it optimizes a loss function by combining multiple weak learners (typically decision trees) to form a strong predictive model.

Table 2. Hyperparameter Tuning for Gradient Boosting

Hyperparameter	Values	Description
n_estimators	randint(50, 100)	Number of boosting stages to run, influencing model accuracy and overfitting. (Aarshay, 2022)
max_depth	randint(3, 10)	Maximum depth of the individual trees, affecting model complexity. (Aarshay, 2022)
min_samples_split	randint(5, 20)	Minimum number of samples required to split an internal node, impacting the model's robustness. (Aarshay, 2022)
min_samples_leaf	randint(5, 20)	Minimum number of samples required at a leaf node, affecting the model's ability to generalize. (Aarshay, 2022)
learning_rate	uniform(0.01, 0.1)	Learning rate, shrinks the contribution of each tree, controlling the pace of learning. (Aarshay, 2022)

Best parameters for Gradient Boosting Classifier --

learning_rate: 0.03539, max_depth: 6 , min_samples_leaf: 18 , min_samples_split: 11, n_estimators: 51
Best cross-validation score for Gradient Boosting: 0.49802555168408824

- Support Vector Machine

SVM is a powerful classifier that works by finding the optimal hyperplane that separates classes in a high-dimensional space. It uses kernels to handle non-linearly separable data.

Table 3. Hyperparameter Tuning for SVM

Hyperparameter	Values	Description
C	uniform(0.1, 10)	Regularization parameter, controlling the trade-off between maximizing the margin and minimizing classification error. (Bala, 2023)
Gamma	['scale', 'auto']	Kernel coefficient, influencing the model's flexibility and decision boundary. (Bala, 2023)
kernel	['linear', 'rbf']	Type of kernel function to be used in the algorithm, defining the hyperplane in higher-dimensional space. (Bala, 2023)

Best parameters for Support Vector Machine –

C: 0.1552, gamma: 'scale' , kernel: 'linear'.

Best cross-validation score for SVM: 0.43484320557491285

An ensemble model, the Voting Classifier, was utilized in this study to combine the predictions of Random Forest, Gradient Boosting, and SVM. This classifier employed a "soft" voting strategy, where it averaged the predicted probabilities from each base model to provide a more balanced and robust final prediction by leveraging the strengths of the individual models.

A voting classifier works by aggregating the predictions from multiple models to determine the final output class based on a majority vote. Instead of using separate, specialized models, a voting classifier integrates their outputs to predict the class with the highest combined votes. There are two types of voting strategies: **Hard Voting**, where the final prediction is the class most frequently predicted by individual models, and **Soft Voting**, where the final prediction is determined by the highest average probability of all models for each class. This approach allows the ensemble model to capitalize on the strengths of multiple classifiers, thereby enhancing overall predictive performance. (geeksforgeek, 2023)

The model evaluation utilized confusion matrices and classification reports to assess precision, recall, F1-score, and accuracy for each class (Short, Medium, Long) across the training, hold-out validation, and external validation sets to gauge generalization to unseen data. A Dummy

Classifier was also used as a baseline, predicting the most frequent class to provide a simple benchmark. By comparing advanced machine learning models against this baseline, it was possible to demonstrate the models' ability to learn meaningful patterns rather than just guessing, thereby validating their robustness and effectiveness. (Tezcan, 2021)

CHAPTER 6

6. RESULTS

6.1 Brain Tumor Classification

6.1.1 Analysis Based on Loss and Accuracy Graphs

- **Inception v3**

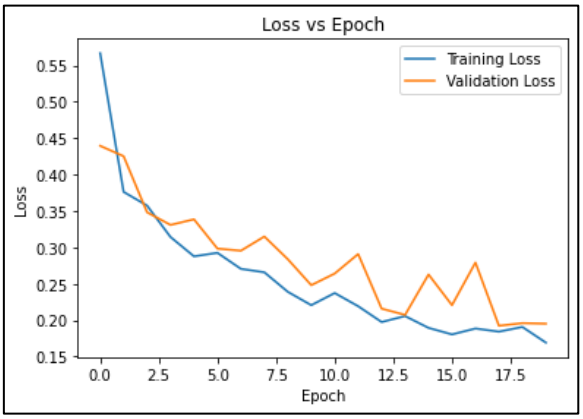


Figure 21. Inception training and validation Loss

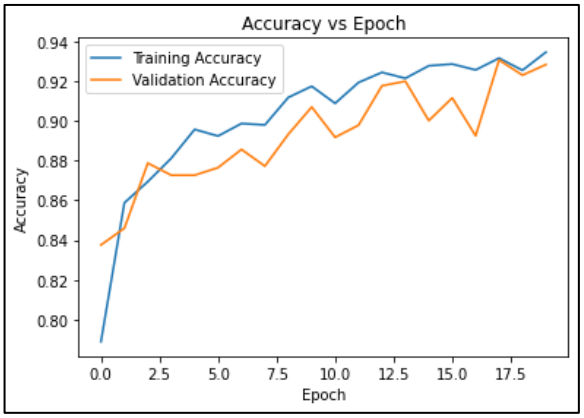


Figure 22. Inception training and validation Accuracy

InceptionV3 demonstrated a consistently high accuracy of around 93-94% over 20 epochs, with both training and validation accuracy closely aligned, indicating effective learning without significant overfitting. The loss curve consistently decreased, further confirming that the model was optimizing effectively during training.

- **MoblienetV2**

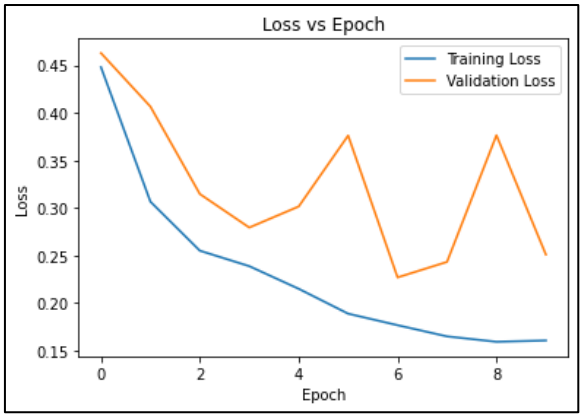


Figure 23. MobileNetV2 training and validation Loss

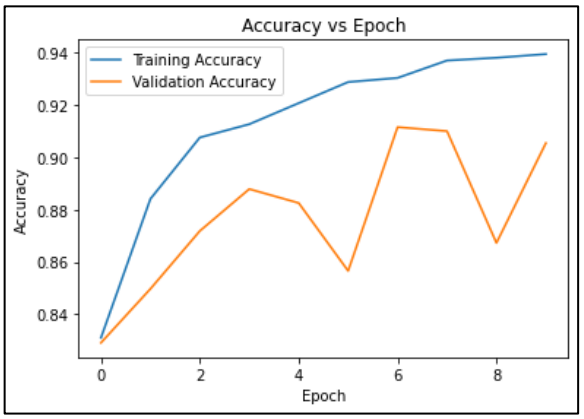


Figure 24. MobileNetV2 training and validation Accuracy

MobileNetV2 also performed well, achieving an overall accuracy of 91% after 10 epochs. The training accuracy nearly reached 94%, but the validation accuracy fluctuated slightly around 90%. The steady decrease in training loss with some variability in validation loss suggests potential overfitting, though the model still maintained strong performance.

- **Resnet152V2**

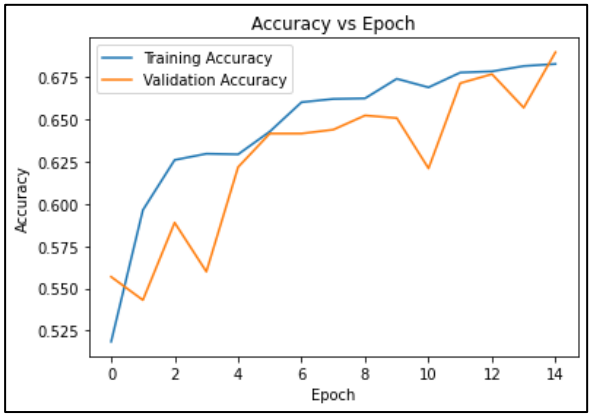


Figure 25. Resnet152V2 training and validation Accuracy

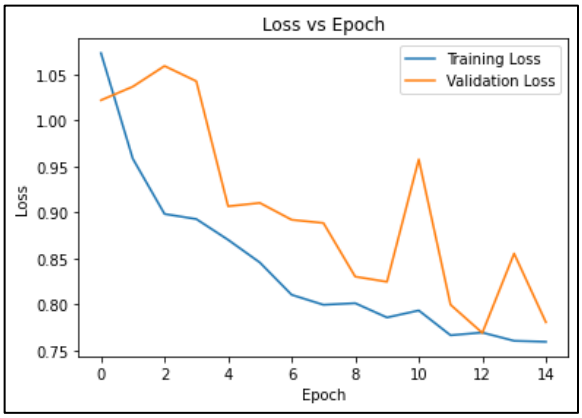


Figure 26. Resnet152V2 training and validation Loss

ResNet152V2 showed a more moderate performance with a final accuracy of around 67%. The training accuracy reached approximately 67.5%, but the validation accuracy plateaued, reflecting limited improvement after a certain point. The fluctuating validation loss further hinted at overfitting or inconsistencies in handling the validation data.

- **VGG19**

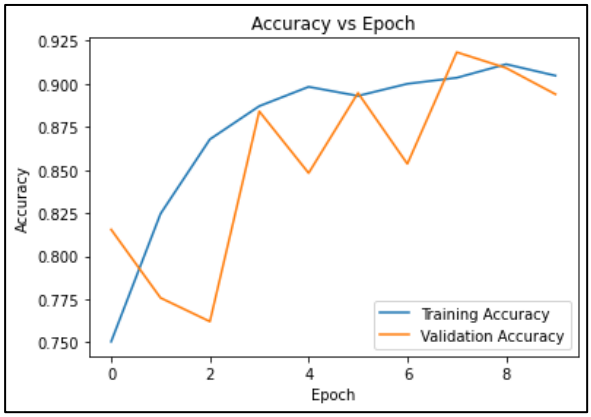


Figure 27. VGG19 training and validation Accuracy

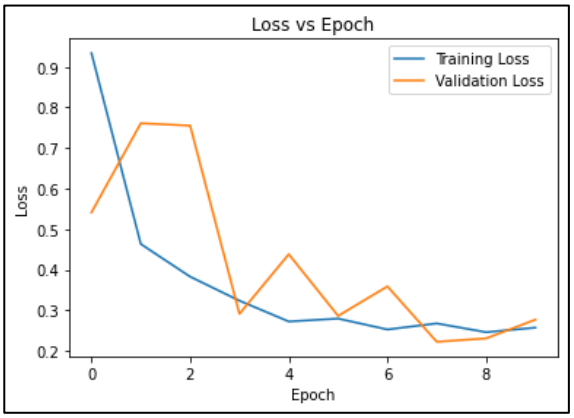


Figure 28. VGG19 training and validation Loss

VGG19 had a relatively high training accuracy, peaking just above 90%, but the validation accuracy showed greater variability, stabilizing around 90%. The decreasing trend in training and validation losses, albeit with some variability, indicates that the model learned effectively, though with some instability in validation performance.

In conclusion, InceptionV3 led in both training and validation accuracy, indicating robust and consistent performance. MobileNetV2 was a close second but showed signs of slight overfitting. ResNet struggled to improve beyond a certain point, and VGG, while strong in training, exhibited some instability in validation accuracy.

6.1.2 Analysis Based on Classification report

- **InceptionV3** excelled in the "notumor" class with near-perfect precision and recall, making it highly reliable for identifying non-tumor cases. The model also performed well in the "glioma" and "meningioma" classes, though precision and recall values were slightly lower, ranging from 84% to 95%, indicating room for improvement in distinguishing between these tumor types.

Classification Report:	precision	recall	f1-score	support
glioma	0.93	0.89	0.91	300
meningioma	0.84	0.88	0.86	306
notumor	0.98	0.99	0.98	405
pituitary	0.95	0.94	0.94	300
accuracy			0.93	1311
macro avg	0.93	0.92	0.92	1311
weighted avg	0.93	0.93	0.93	1311

Figure 29. Classification report of InceptionV3

- **MobileNetV2** showed the highest precision and recall for the "notumor" class, nearly perfect at 98% and 99%. However, its performance in the "glioma" and "meningioma" classes was slightly lower, with recall scores of 79% and 82%, respectively, indicating that the model sometimes struggled to correctly identify these tumor types. The pituitary class performed well, with a recall of 99%, indicating strong reliability in detecting this tumor type.

Classification Report:	precision	recall	f1-score	support
glioma	0.95	0.79	0.86	300
meningioma	0.82	0.82	0.82	306
notumor	0.98	0.99	0.99	405
pituitary	0.86	0.99	0.92	300
accuracy			0.91	1311
macro avg	0.90	0.90	0.90	1311
weighted avg	0.91	0.91	0.90	1311

Figure 30. Classification report of MobilenetV2

- **ResNet** had the highest F1-score of 0.86 for the "no tumor" class, reflecting strong performance in this category. However, the "glioma" class exhibits lower recall at 0.39, indicating difficulties in correctly identifying all glioma instances. The overall accuracy of 69% suggests that while the model is performing moderately well, there is room for improvement, particularly in handling classes with more complex features or fewer training samples.

Classification Report:	precision	recall	f1-score	support
glioma	0.81	0.39	0.53	300
meningioma	0.53	0.56	0.54	306
notumor	0.82	0.90	0.86	405
pituitary	0.63	0.85	0.72	300
accuracy			0.69	1311
macro avg	0.70	0.67	0.66	1311
weighted avg	0.71	0.69	0.68	1311

Figure 31. Classification report of Resnet152V2

- **VGG** performed exceptionally well in the "notumor" class, with precision, recall, and F1-scores near 97%. The "glioma" and "pituitary" classes were also identified with high accuracy, with precision scores of 94% and 84%, respectively. However, the "meningioma" class had slightly lower precision and recall, indicating a potential area where the model could be improved.

Classification Report:	precision	recall	f1-score	support
glioma	0.94	0.78	0.85	300
meningioma	0.80	0.81	0.81	306
notumor	0.98	0.97	0.97	405
pituitary	0.84	1.00	0.91	300
accuracy			0.89	1311
macro avg	0.89	0.89	0.89	1311
weighted avg	0.90	0.89	0.89	1311

Figure 32. Classification report of VGG19

In conclusion, InceptionV3 and MobileNetV2 were the most reliable in identifying "notumor" cases, with InceptionV3 leading slightly in overall balance across all classes. ResNet struggled significantly with glioma classification, and while VGG performed well overall, it also faced challenges in differentiating between similar tumor types.

6.1.3 Analysis Based on Confusion Matrix

- **InceptionV3**

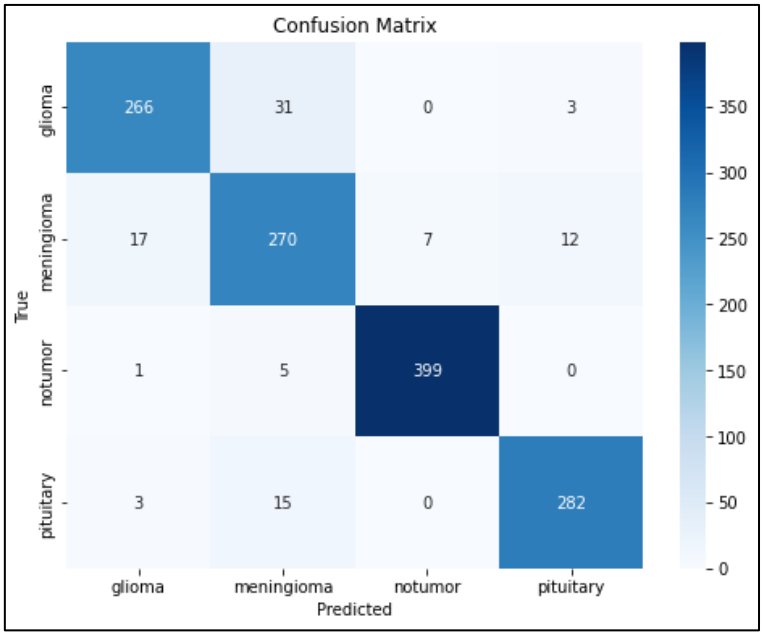


Figure 33. Confusion Matrix of InceptionV3

The confusion matrix revealed that the primary source of misclassification was between glioma and meningioma, which may be due to the visual similarities between these tumor types in MRI images. Despite this, the model was highly accurate in classifying "notumor" and "pituitary" cases. The strong performance across most classes suggests that the InceptionV3 model is a robust tool for brain tumor classification, though targeted improvements could enhance its ability to differentiate between more similar classes.

- **MobileNetV2**

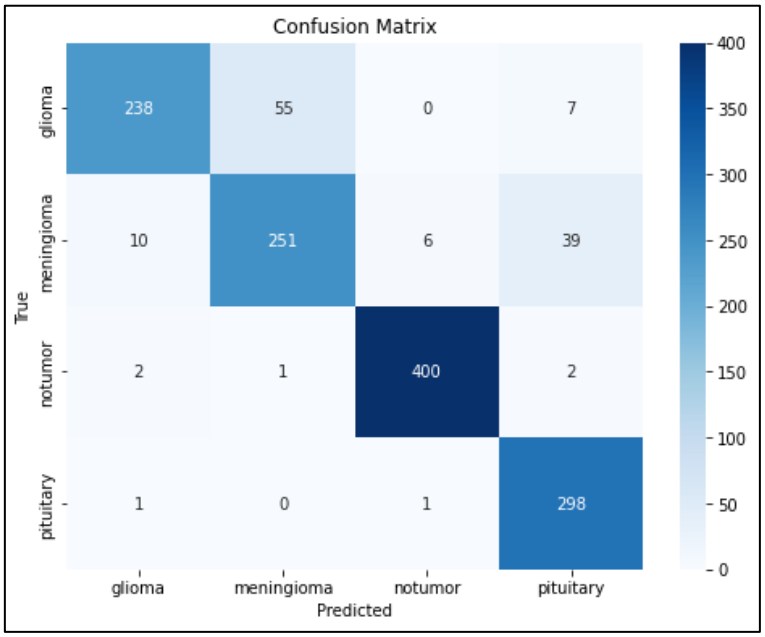


Figure 34. Confusion Matrix of MobileNetV2

The confusion matrix highlights that most misclassifications occurred between the glioma and meningioma classes, similar to the InceptionV3 model. Despite these challenges, the MobileNetV2 model remains a strong performer, particularly in identifying non-tumor cases and pituitary tumors. The results suggest that while MobileNetV2 is effective, further tuning or additional data might be necessary to improve differentiation between more visually similar tumor types.

- **Resnet152V2**

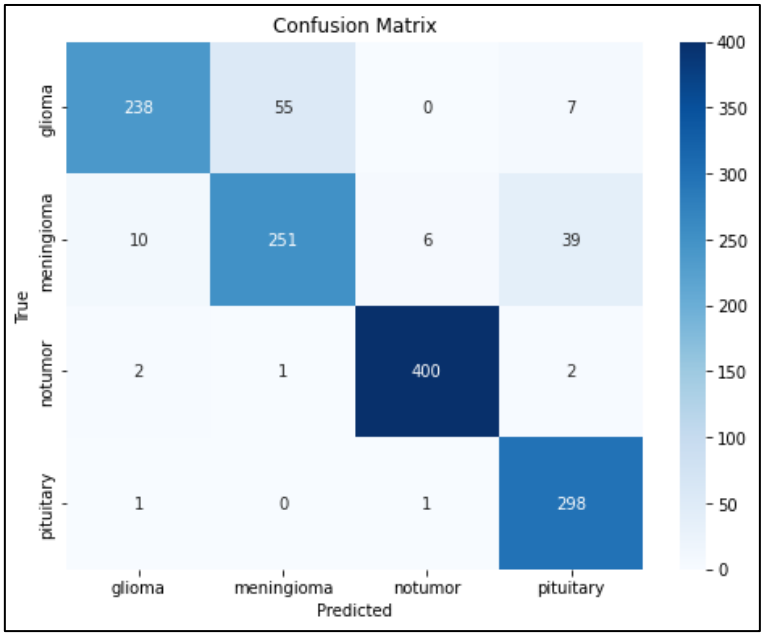


Figure 35. Confusion matrix of Resnet152V2

The confusion matrix reveals specific areas where the model struggles. For instance, there is significant misclassification between the "glioma" and "meningioma" classes, with a substantial number of glioma images being incorrectly labeled as meningioma. This confusion could stem from the visual similarities between these tumor types, highlighting a potential area for further refinement in the model or data augmentation to improve differentiation.

- **VGG19**

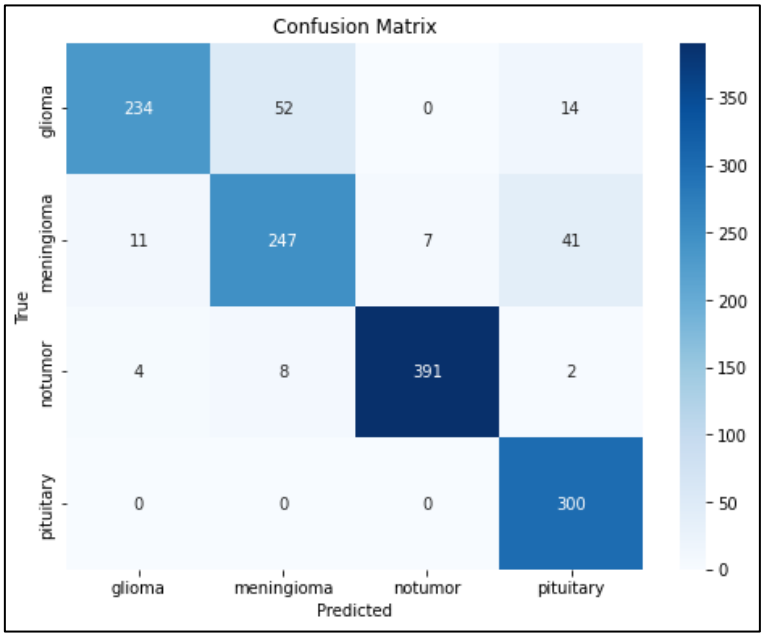


Figure 36. Confusion matrix of VGG19

The confusion matrix further illustrates the VGG model's classification performance, showing that most errors occurred when differentiating between the "glioma" and "meningioma" classes. Despite these challenges, the model consistently identified the "notumor" and "pituitary" classes with high precision and recall, contributing to an overall accuracy of 89%. This evaluation indicates that while the VGG model performs well, particularly in recognizing certain classes, further optimization could enhance its performance across all categories.

6.1.4 Ensemble Techniques Evaluation

Table 4. Class specific Performance of Ensemble Techniques

Class	Metric	Simple Averaging	Geometric Mean	Weighted Averaging
Class 0 (Glioma)	Precision	0.98	0.98	0.98
	Recall	0.80	0.80	0.81
Class 1 (Meningioma)	Precision	0.83	0.82	0.83
	Recall	0.89	0.90	0.89
Class 2 (No Tumor)	Precision	0.99	0.99	0.99
	Recall	1.00	0.99	1.00
Class 3 (Pituitary)	Precision	0.91	0.92	0.91
	Recall	1.00	1.00	1.00

The above table presents a comparison of precision and recall metrics across three ensemble methods—simple averaging, geometric mean, and weighted averaging—applied to the classification of brain tumors. The metrics for Class 2 ("No Tumor") and Class 3 ("Pituitary") show near-perfect precision and recall, with values of 0.99 and 1.00, respectively, across all methods, indicating the models' strong performance in identifying these classes. However, the perfect or near-perfect recall, particularly for Class 2 and Class 3, suggests a risk of overfitting, where the models may be too finely tuned to the training data, potentially reducing their generalization capability on unseen data. For Class 0 ("Glioma") and Class 1 ("Meningioma"), the precision and recall are slightly lower, with minor variations across the ensemble methods, indicating consistent but slightly less confident performance in these categories. This suggests that while the ensemble methods effectively classify most classes, caution should be taken to ensure the models generalizability, especially for the classes with perfect recall.

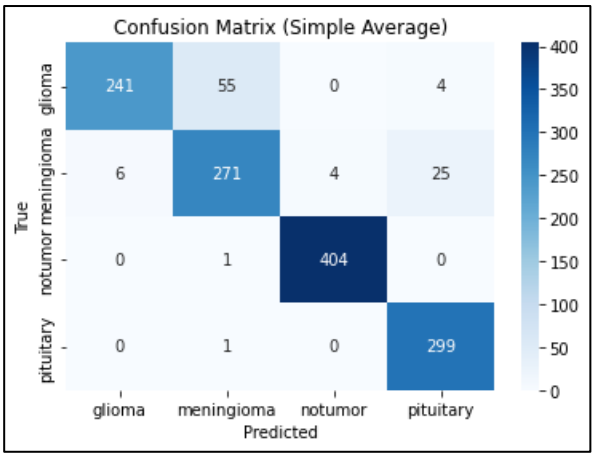


Figure 37. Confusion matrix of simple average

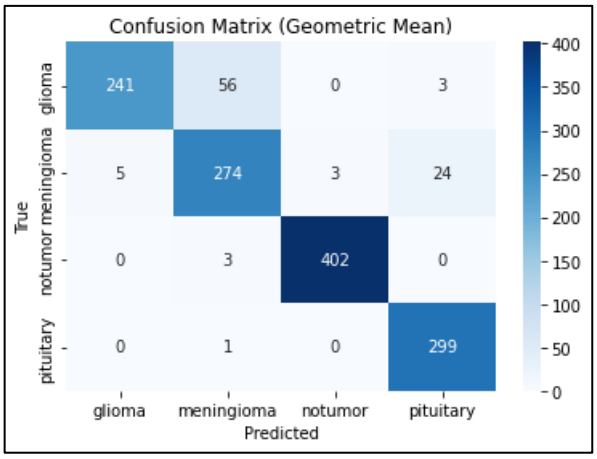


Figure 38. Confusion matrix of geometric Mean

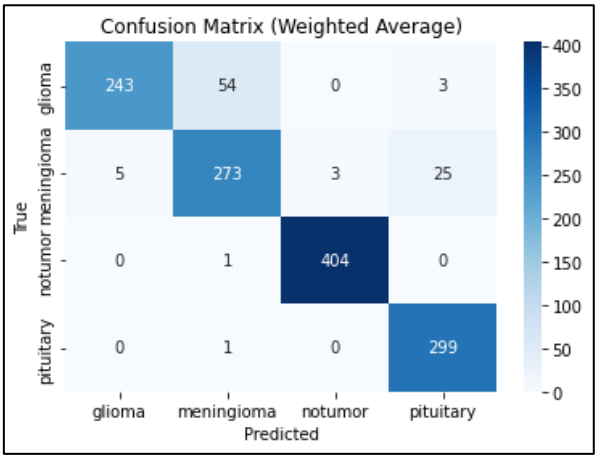


Figure 39. Confusion matrix of weighted Average

The confusion matrices provide further insights into the ensemble methods' performance. Although these methods generally reduced the number of misclassifications, they also revealed specific challenges. Misclassifications were most frequent between the 'glioma' and 'meningioma' classes, suggesting that these tumor types have overlapping features that the models struggle to differentiate. Nonetheless, the ensemble techniques demonstrated a strong ability to accurately classify the 'notumor' and 'pituitary' classes, reinforcing their overall effectiveness. These findings highlight the strengths of ensemble methods in enhancing classification performance, while also pointing out areas, such as distinguishing between certain tumor types, where further model refinement may be needed.

6.1.5 Comparison with Baseline Model for brain tumor classification

Dummy Classifier Performance:

	precision	recall	f1-score	support
0	0.00	0.00	0.00	300
1	0.00	0.00	0.00	306
2	0.31	1.00	0.47	405
3	0.00	0.00	0.00	300
accuracy			0.31	1311
macro avg	0.08	0.25	0.12	1311
weighted avg	0.10	0.31	0.15	1311

Figure 40. Classification report of dummy classifer for tumor classification

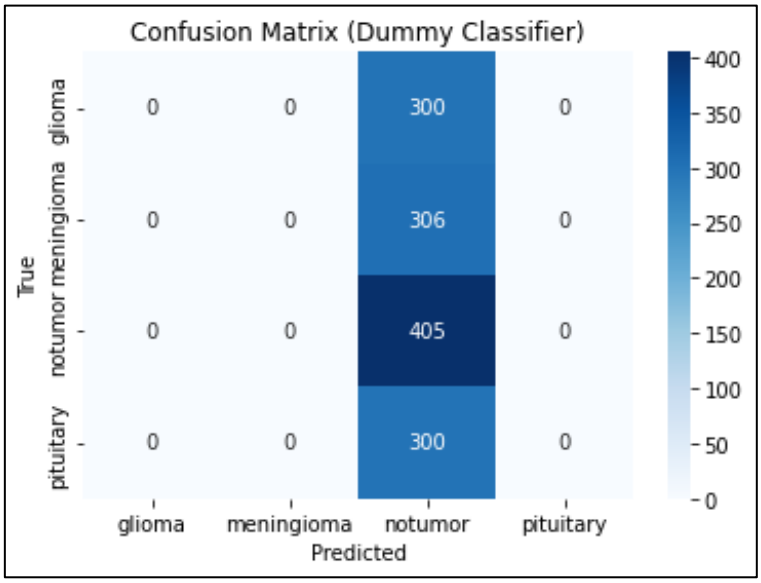


Figure 41. Confusion Matrix of dummy classifier for tumor classification

For a baseline comparison, a dummy classifier was used, which predicted the most frequent class (no tumor) for all instances. As expected, the dummy classifier performed poorly, achieving an accuracy of only 31%, with a precision, recall, and F1-score of 0.00 for all classes except the most frequent one. This stark contrast highlights the effectiveness of the transfer learning models and ensemble methods, which significantly outperformed this naive approach.

6.1.6 Performance Trade-Off Analysis

In addition to accuracy and precision, inference time was also considered to evaluate the trade-off between model performance and computational efficiency. The trade-off between model performance and computational efficiency was analysed using inference time and the Matthews Correlation Coefficient (MCC). Inference time measures how long a model takes to make

predictions, crucial for real-time applications. MCC evaluates a model's predictive performance, especially in imbalanced datasets, by considering all confusion matrix elements.

The trade-off scores were calculated by considering the ideal model with perfect MCC and minimal inference time. The analysis showed that ensemble methods (Average, Weighted Average, Geometric Mean) achieved high MCC scores (around 0.90) but had significantly high inference times (13.69159), making them less practical for real-time use.

Among individual models, **MobileNetV2** provided the best balance, with an MCC of 0.8746 and the lowest inference time of 0.63146, resulting in a favourable trade-off score of 0.6437. This indicates that MobileNetV2 is the most suitable for applications requiring both accuracy and efficiency. In contrast, models like ResNet152 and VGG19, despite good MCC scores (0.5907 and 0.8598), had longer inference times, resulting in higher trade-off scores (5.0410 and 6.9251). This analysis highlights the need to consider both accuracy and computational cost when selecting models for deployment.

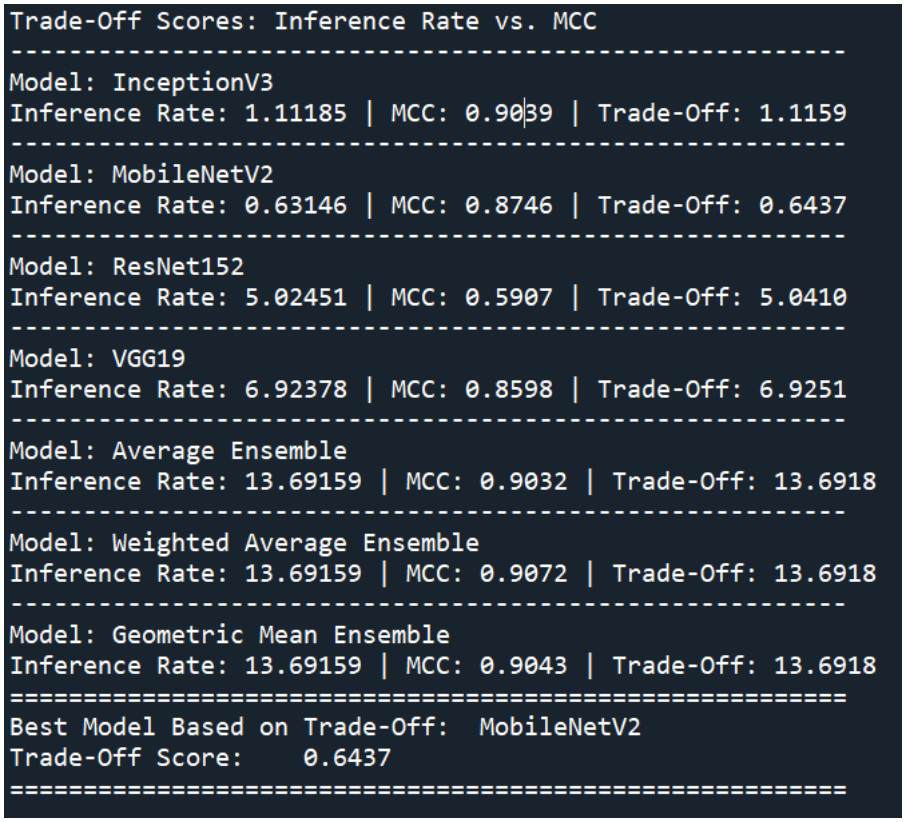


Figure 42. Trade off scores for each model

6.2 Brain tumor Segmentation and Survival Prediction Results

6.2.1 Segmentation Results

The trained model was subsequently evaluated on a separate validation set, where predictions were generated for the MRI scans. These predictions were compared against the ground truth segmentation masks, and the Mean IoU was calculated to quantify the model's performance. The model achieved promising results, indicating its ability to accurately segment brain tumors from MRI scans.

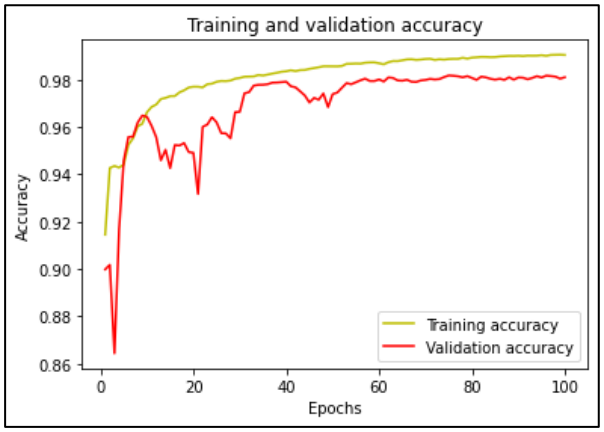


Figure 43. Accuracy for 3D U-Net

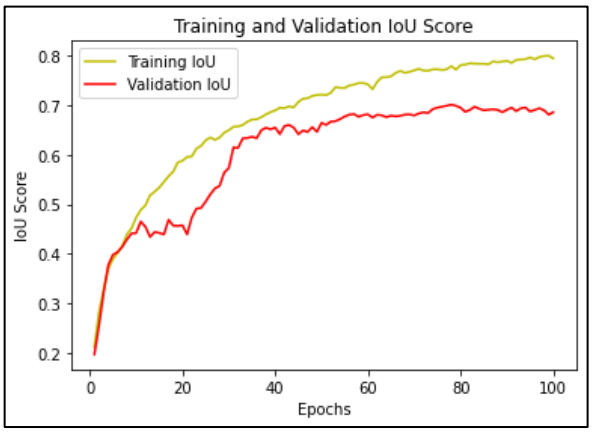


Figure 44. IOU Score for 3D U-Net

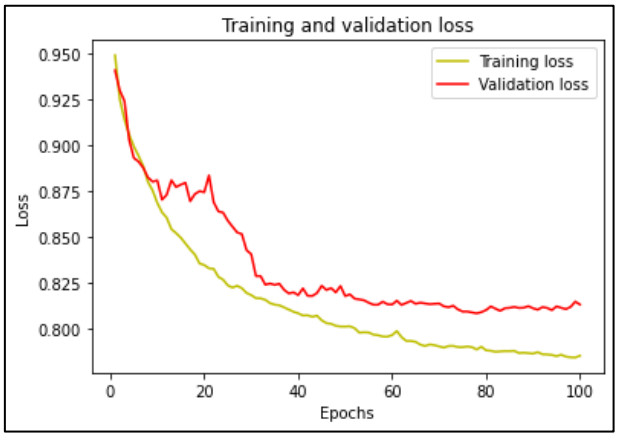


Figure 45. Loss for 3D U-Net

Figure 43, **Figure 44** & **Figure 45** shows, **the 3D U-Net model** showed consistent improvement in both accuracy and IoU scores over the training epochs, indicating that it was effectively learning to segment the different tumor regions. However, a subsequent fine-tuning of the 3D U-net model for an additional 10 epochs, shown in **Figure 46** & **Figure 47** did not

result in significant improvements, suggesting that the model had already reached its optimal performance.

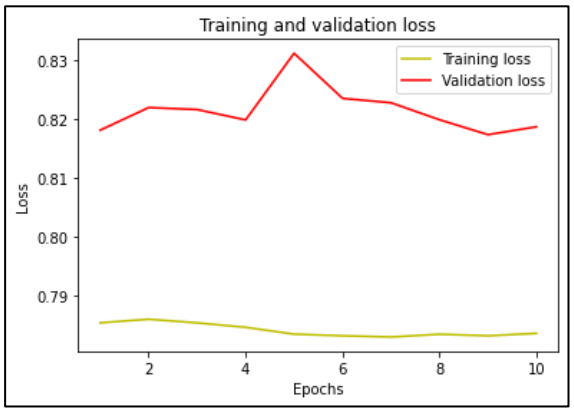


Figure 46. Loss for 3D U-Net (10epochs)

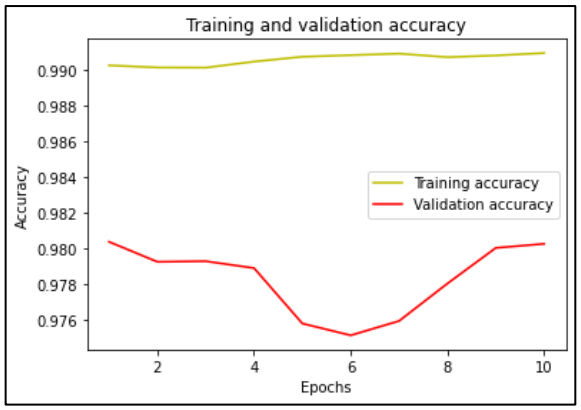


Figure 47. Accuracy for 3D U-Net (10 epochs)

The attention U-net training history in **Figure 48** revealed a consistent improvement in both accuracy and IoU over the training epochs, suggesting effective learning as well. The training loss decreases steadily, with a sharp drop mid-training, while the validation loss fluctuates, spiking around epoch 60, indicating potential overfitting that later corrects itself. Training accuracy improves consistently, nearing perfection, while validation accuracy shows more

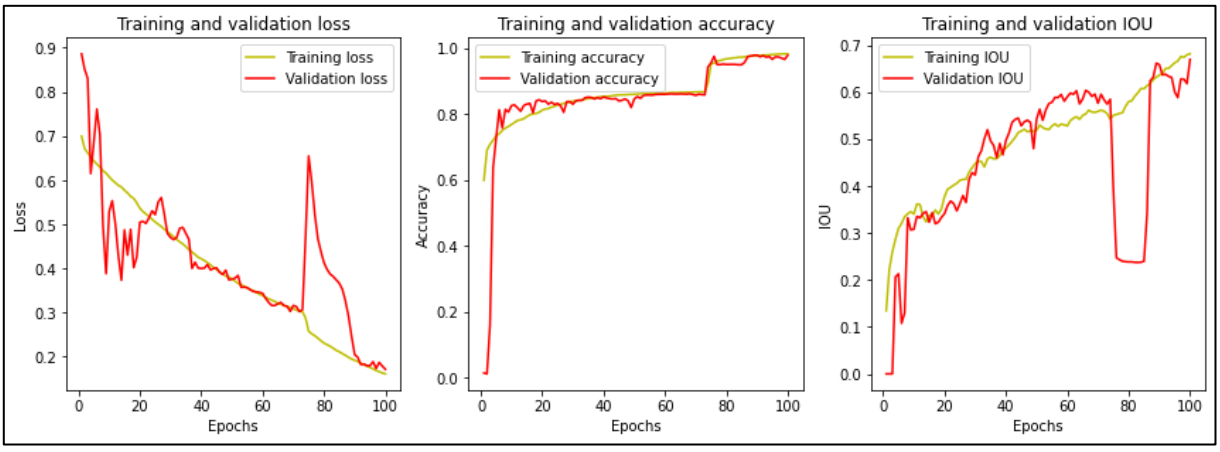


Figure 48. Accuracy, Loss & IoU Score for 3D Attention U-Net

fluctuation due to the model's complexity. The IoU follows a similar pattern, with validation IoU dipping around epoch 60 before recovering, reflecting the model's correction after a phase of instability.

In conclusion, the standard U-Net showed more stable training and validation curves, while the Attention U-Net exhibited some variability, possibly due to the added complexity of the attention mechanisms. The 3D Attention U-Net, while potentially more powerful due to its attention mechanisms, requires more careful tuning and may be more sensitive to changes in training data or hyperparameters, as indicated by the more erratic validation curves. In terms of final performance metrics (accuracy and IoU), both models achieve high scores, but the 3D U-Net appears to reach these scores with more consistency, while the 3D Attention U-Net shows more potential variability.

Table 5. Comparison of Mean IoU Scores 3D U-NET & 3D Attention U-Net

	3D U-Net Model	**3D Attention U-Net Model**
Mean IoU (validation)	0.7454	0.6919

The 3D U-Net outperformed the 3D Attention U-Net in terms of overall IoU, achieving a higher score of 0.7454 compared to 0.6919 as shown in **Table 5**. This suggests that the base U-Net architecture was slightly more effective in generalizing to unseen data.

Table 6. Class-Wise Accuracy Comparison Between 3D U-Net & 3D Attention U-Net

Accuracy Class-Wise	**3D U-Net Model**	**3D Attention U-Net Model**
Class 0 (Background)	99.66%	98.82%
Class 1 (Core)	64.04%	62.30%
Class 2 (Edema)	84.55%	92.95%
Class 3 (Enhancing Tumor)	80.67%	83.64%

The class-wise accuracy results for the 3D U-Net and 3D Attention U-Net models in **Table 6** show some interesting contrasts:

1. Class 0 (Background): Both models achieve high accuracy in identifying background regions, with the 3D U-Net slightly outperforming the 3D Attention U-Net. This result suggests that both models are very effective at distinguishing non-tumor areas, though the 3D U-Net has a marginal advantage.
2. Class 1 (Core): The accuracy for identifying necrotic core areas is lower for both models, with the 3D U-Net slightly outperforming the 3D Attention U-Net. This

indicates that core tissue is challenging to segment accurately, likely due to its similarity to surrounding tumor regions, and the attention mechanism did not significantly improve performance for this class.

3. Class 2 (Edema): The 3D Attention U-Net shows a significant improvement over the 3D U-Net in identifying edema, achieving a higher accuracy. This suggests that the attention mechanism is particularly effective in enhancing the model's focus on diffuse and possibly less distinct tumor boundaries, which are characteristic of edema.
4. Class 3 (Enhancing Tumor): The 3D Attention U-Net also outperforms the 3D U-Net in identifying enhancing tumor regions. This improvement highlights the attention mechanism's capability to better capture and focus on the critical areas of enhancing tumor tissue, which are essential for accurate tumor characterization and treatment planning.

Overall, the 3D Attention U-Net demonstrates a clear advantage in identifying edema and enhancing tumor regions, likely due to its enhanced focus provided by the attention mechanisms. However, for simpler tasks such as identifying background and core areas, the 3D U-Net performs slightly better or equally well, suggesting that the attention mechanism's benefits are most pronounced in more complex segmentation tasks.

Furthermore, visualizations of the predictions were generated to qualitatively assess the model's performance. The input images, ground truth masks, and predicted masks were plotted side by side, providing a clear visual representation of both the model's segmentation capabilities.

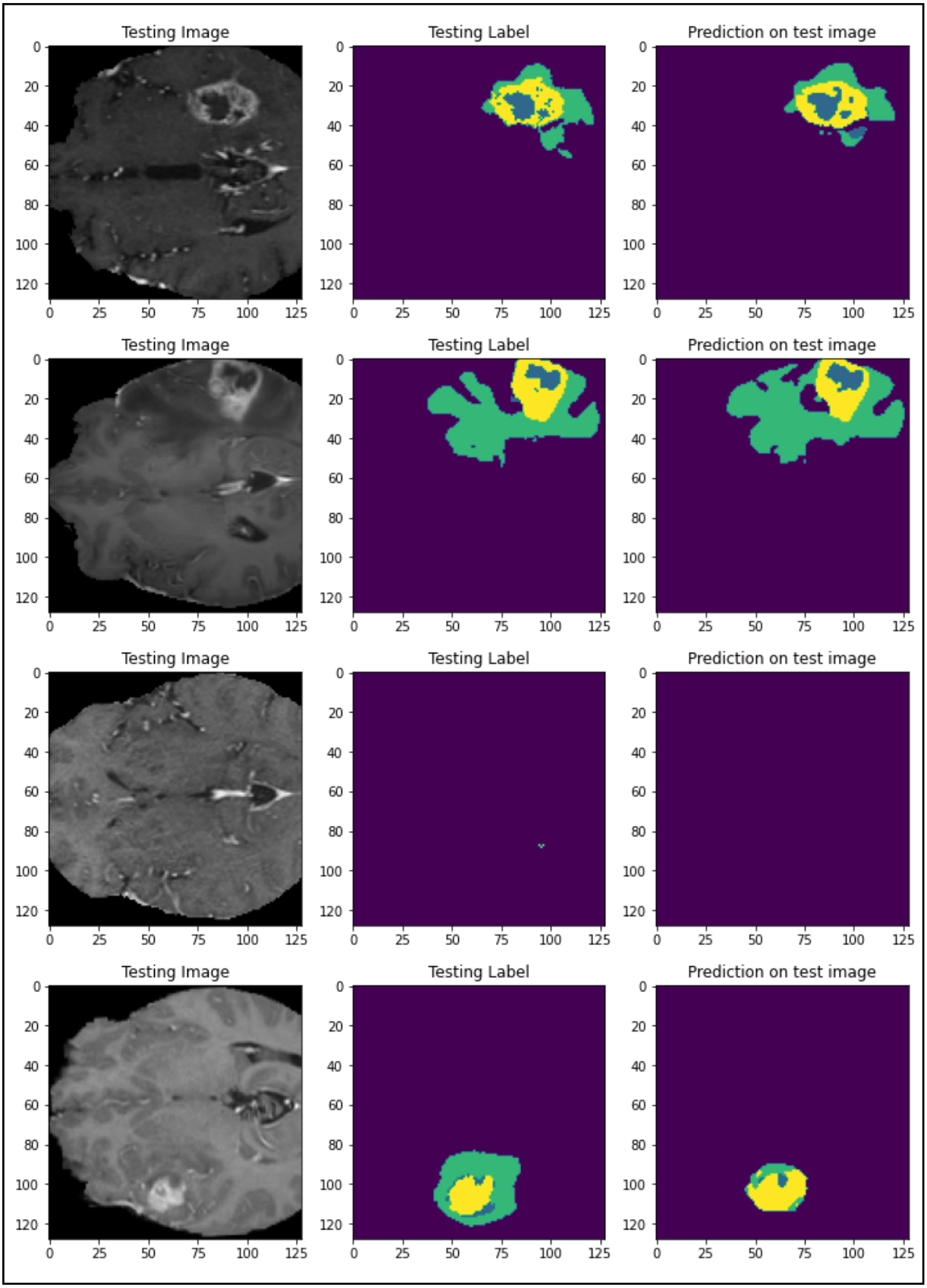


Figure 49. Testing Images and Labels vs Predicted Labels for 3D U-Net

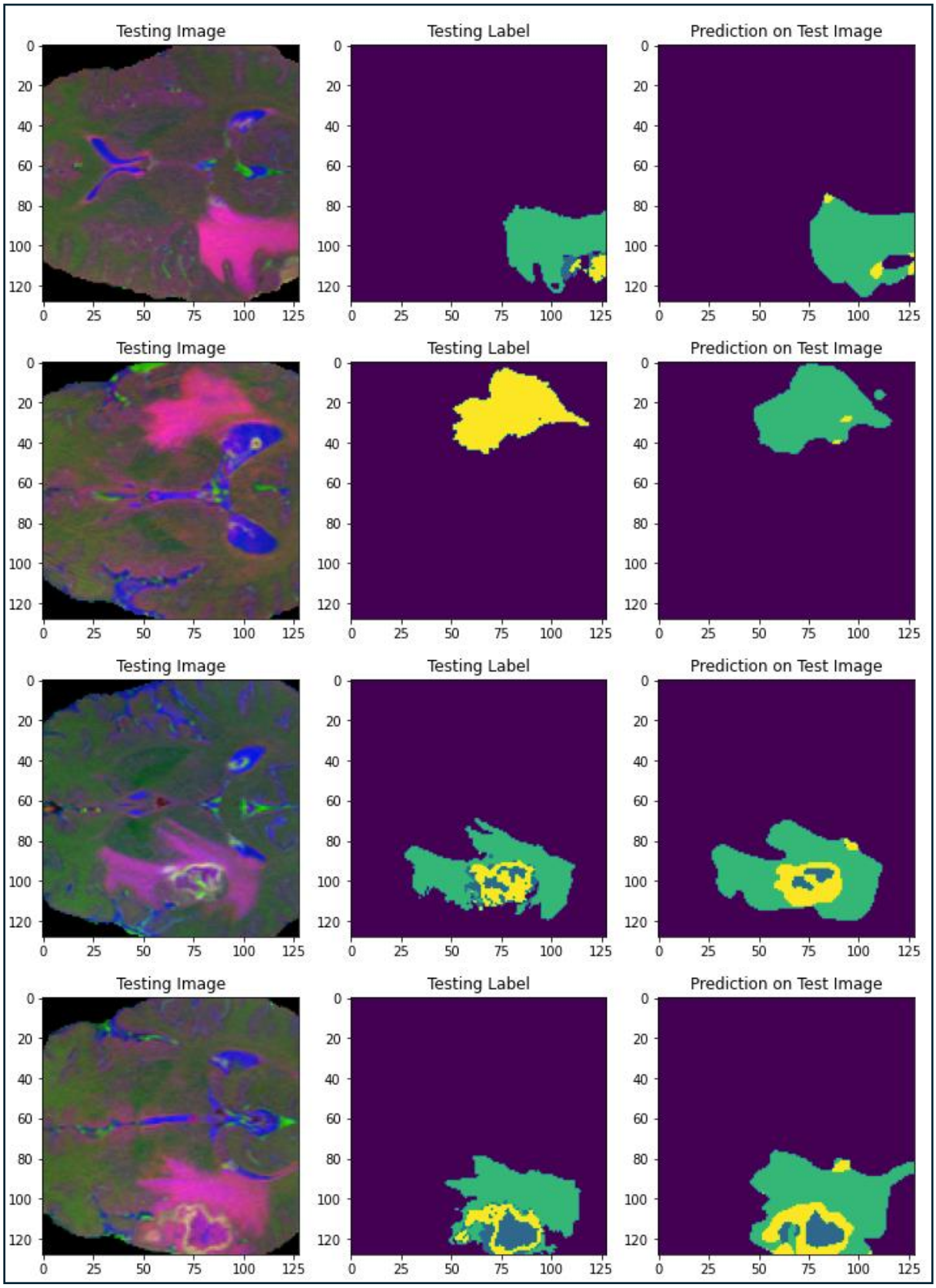


Figure 50. Testing Images and Labels vs Predicted Images for 3D Attention U-Net

The colors represent specific regions –

- green for edema,
- yellow for the core, and
- blue for the enhancing tumor.

For the 3D U-Net model, the visualizations of results in **Figure 49** & **Figure 50** reveal that the model accurately segments the major regions of the tumors. The green regions (edema) are generally well-captured, indicating that the model is effectively identifying areas of swelling around the tumor. The yellow regions (core) and blue regions (enhancing tumor) are also identified, although there might be some instances where the boundaries are not as sharp as they could be. This suggests that while the 3D U-Net is proficient in segmenting larger and more distinct regions like the edema, it may struggle slightly with finer details in the necrotic core and enhancing tumor regions.

In the case of the 3D Attention U-Net, the results show an improvement in capturing smaller and more complex regions. The attention mechanism enhances the model's ability to focus on the blue (enhancing tumor) and yellow (core) regions, which are crucial for accurate tumor characterization. The green (edema) regions are still well-defined, but the attention model's strength lies in its improved handling of the necrotic core and enhancing tumor areas, where it provides more precise and sharper segmentation compared to the 3D U-Net.

In summary, the 3D Attention U-Net demonstrates a superior capability in delineating the necrotic core and enhancing tumor regions, making it a more suitable model for tasks where detailed segmentation of these critical areas is necessary. However, the 3D U-Net remains a robust model for broader segmentation tasks, especially where the focus is on larger regions like edema.

6.2.2 Survival Prediction Results

Table 7. Classification Report Table for Survival Prediction

Accuracy	Random Forest	Gradient Boosting	SVM	Voting Classifier
Training	0.77	0.76	0.46	0.74
Validation	0.48	0.52	0.48	0.56

Classifier	Dataset	Class	Precision	Recall	F1-Score
Random Forest	Training Set	Short	0.76	0.83	0.79
		Medium	0.81	0.74	0.77
		Long	0.74	0.74	0.74
	Validation Set	Short	0.67	0.40	0.50
		Medium	0.32	0.40	0.35
		Long	0.55	0.61	0.58
Gradient Boosting	Training Set	Short	0.77	0.81	0.79
		Medium	0.80	0.71	0.75
		Long	0.73	0.77	0.75
	Validation Set	Short	0.58	0.47	0.52
		Medium	0.44	0.47	0.45
		Long	0.55	0.61	0.58
SVM	Training Set	Short	0.49	0.61	0.54
		Medium	0.39	0.17	0.24
		Long	0.46	0.59	0.52
	Validation Set	Short	0.78	0.47	0.58
		Medium	0.18	0.13	0.15
		Long	0.50	0.78	0.61
Voting Classifier	Training Set	Short	0.72	0.722	0.74
		Medium	0.80	0.68	0.73
		Long	0.72	0.77	0.74
	Validation Set	Short	0.70	0.47	0.56
		Medium	0.41	0.47	0.44
		Long	0.62	0.72	0.67

The **Table 7** summarizes the performance of four models—Gradient Boosting, Random Forest, SVM, and Voting Classifier—on both training and validation datasets for predicting survival durations categorized as Short, Medium, or Long.

For the **Random Forest Classifier**, the training set accuracy was 77%, with strong F1-scores across all classes (Short: 0.79, Medium: 0.77, Long: 0.74). However, the validation set accuracy dropped significantly to 48%, with particularly low F1-scores for the "Medium" class (0.35). This drop indicates that while Random Forest performed well on the training data, it struggled to generalize to unseen data, especially for the "Medium" class.

The **Gradient Boosting Classifier** achieved a training set accuracy of 76%, showing balanced F1-scores for all classes (around 0.75-0.79). On the validation set, the accuracy decreased to 52%, with the F1-score for the "Short" class at 0.52 and for the "Medium" class at 0.45, suggesting moderate generalization capability, but still limited in handling "Medium" cases effectively.

The **Support Vector Machine (SVM)** model showed the lowest training accuracy at 46%, with poor F1-scores, particularly for the "Medium" class (0.24). On the validation set, the SVM achieved an accuracy of 48%. It performed reasonably well for the "Long" class (F1-score: 0.61) but showed weak generalization for the "Medium" class (F1-score: 0.15). This suggests that SVM struggled overall, especially with the "Medium" survival class.

The **Voting Classifier**, combining Random Forest, Gradient Boosting, and SVM, demonstrated a training accuracy of 74% with balanced F1-scores across all classes (around 0.73-0.74). On the validation set, the Voting Classifier outperformed the individual models with an accuracy of 56% and higher F1-scores for all classes, particularly the "Long" class (0.67). This indicates that the Voting Classifier provided more robust and balanced predictions by leveraging the strengths of its component models.

The overfitting observed in all models is evident from the substantial drop in performance from the training to validation datasets, indicating that the models have learned patterns specific to the training data rather than generalizable ones. This overfitting likely results from the simplicity of the dataset, which only includes features like age and tumor volumes without capturing the full complexity of factors affecting survival. The limited feature set leads to high

variance and poor generalization. Enhancing the dataset with additional relevant features or more advanced feature engineering could improve the models' predictive accuracy and ability to generalize to unseen data.

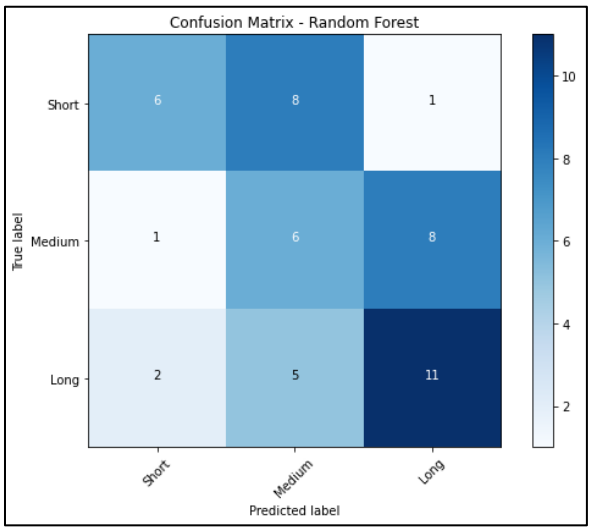


Figure 53. Confusion Matrix for Random Forest

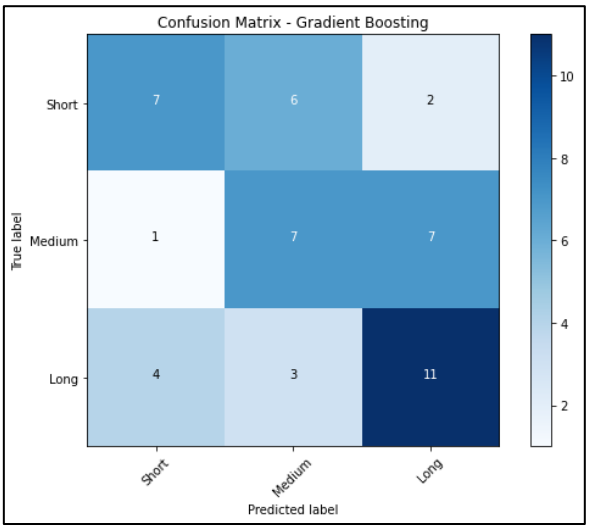


Figure 54. Confusion Matrix for Gradient Boosting

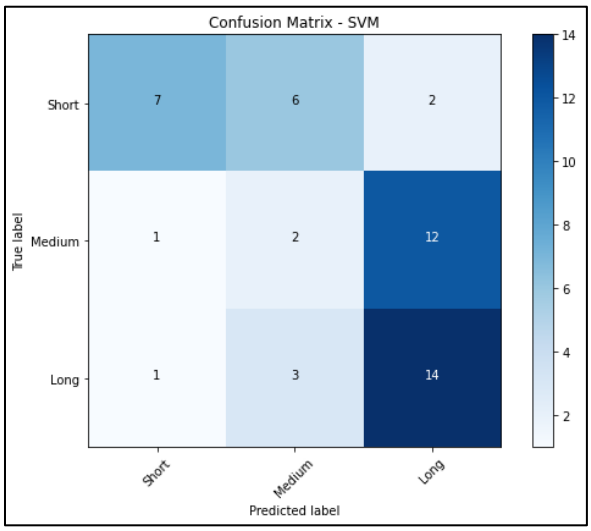


Figure 52. Confusion Matrix for SVM

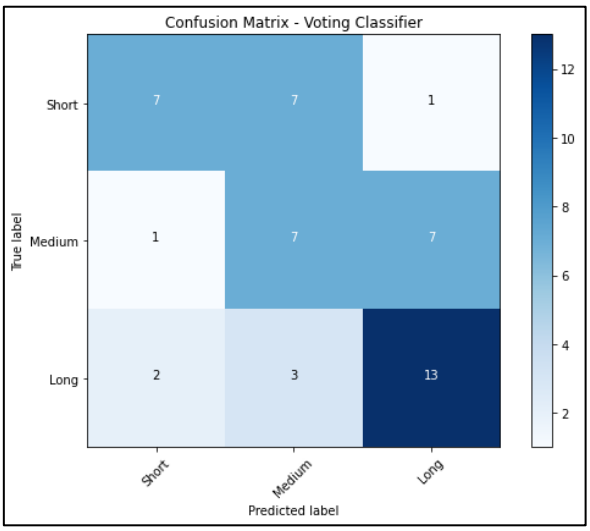


Figure 51. Confusion Matrix for Voting Classifier

The confusion matrices for Gradient Boosting, Random Forest, SVM, and Voting Classifier models reveal various patterns in model performance for the classification task.

Gradient Boosting and Random Forest struggle most with "Medium" survival predictions, often confusing them with "Short" or "Long". SVM has the highest confusion for the "Medium" class, frequently predicting them as "Long," but performs relatively well for "Long" predictions.

Voting Classifier provides the most balanced performance, reducing errors across all classes, but still shows some confusion between adjacent survival categories ("Short" vs. "Medium" and "Medium" vs. "Long").

These results suggest that while ensemble methods like the Voting Classifier can reduce misclassification, individual models have specific strengths and weaknesses that influence their overall performance.

Comparison with Dummy Classifier:

Classification Report for Dummy Classifier (Hold-out Validation Set):

	precision	recall	f1-score	support
Short	0.31	1.00	0.48	15
Medium	0.00	0.00	0.00	15
Long	0.00	0.00	0.00	18
accuracy			0.31	48
macro avg	0.10	0.33	0.16	48
weighted avg	0.10	0.31	0.15	48

Figure 55. Classification Report of Dummy Classifier for Survival Prediction

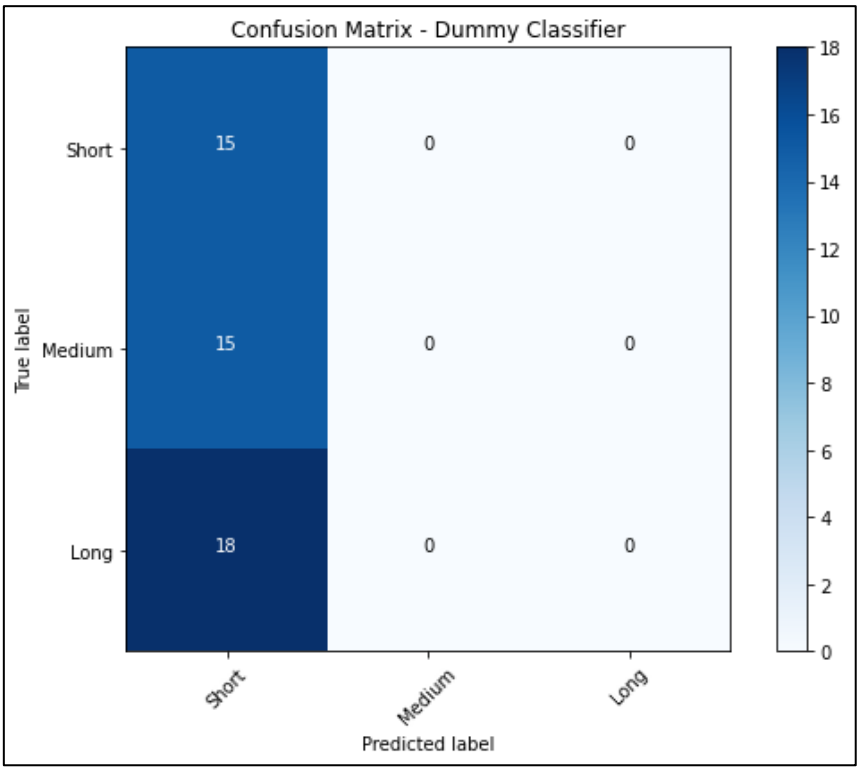


Figure 56. Confusion Matrix of Dummy Classifier for Survival Prediction

The Dummy Classifier, used as a baseline, achieved an accuracy of 31% by predicting the most frequent class, "Short," for all instances. The confusion matrix **Figure 56** shows that it correctly classified all "Short" survival cases but failed to identify any instances of "Medium" or "Long" survival. The precision, recall, and F1-scores for the "Medium" and "Long" categories are zero, highlighting the Dummy Classifier's ineffectiveness in making meaningful predictions. The Dummy Classifier's performance underscores the need for more sophisticated models to achieve any reasonable level of predictive accuracy in this task.

8. CONCLUSION

10.1 Overview

This study explored the application of advanced machine learning and deep learning techniques for brain tumor detection, classification, segmentation, and survival prediction using MRI data. The research employed a range of pre-trained models and innovative architectures, highlighting their respective strengths and challenges. By integrating both classification and segmentation tasks, the study aimed to improve diagnostic accuracy and provide predictive insights that can enhance clinical decision-making in neuro-oncology.

10.2 Summary of the Investigation Study

The investigation involved a comparative analysis of several transfer learning models—InceptionV3, MobileNetV2, ResNet152V2, and VGG19—for brain tumor classification, along with the development of 3D U-Net and 3D Attention U-Net architectures for segmentation tasks. For survival prediction, traditional machine learning models, including Random Forest, Gradient Boosting, and SVM, as well as a Voting Classifier ensemble, were evaluated. The study's multi-faceted approach allowed for a comprehensive assessment of both established and novel techniques in the context of brain tumor analysis.

10.3 Findings and Recommendations

- **Classification Models:**
  - InceptionV3 achieved the highest accuracy (92.83%), with the weighted average ensemble slightly surpassing it at 92.98%. This highlights the effectiveness of ensemble methods in improving model performance.
  - ResNet152V2, despite its complexity, underperformed with an accuracy of 68.95%, indicating that deeper architectures are not always optimal for medical imaging tasks without sufficient data or appropriate tuning.
  - MobileNetV2 provided a good balance between accuracy (90.54%) and computational efficiency, making it suitable for real-time applications.

 Segmentation Models:

- The 3D U-Net model outperformed the 3D Attention U-Net in terms of overall mean IoU (74.54% vs. 69.19%), suggesting that while attention mechanisms can enhance focus on complex tumor regions, they may also introduce variability that requires careful management.
- The 3D Attention U-Net excelled in segmenting smaller and complex regions such as edema and enhancing tumors, showcasing its potential for tasks requiring detailed segmentation.

 Survival Prediction Models:

- Random Forest achieved the highest training accuracy (71%) but exhibited overfitting with a validation accuracy of 56%, underscoring the need for more diverse features beyond the basic clinical data used.
- The Voting Classifier ensemble balanced performance with a consistent validation accuracy of 56%, demonstrating the potential of ensemble methods in mitigating overfitting and improving model robustness.

Based on these findings, the following recommendations are proposed:

Model Optimization: Further fine-tuning and optimization of attention mechanisms could improve the performance consistency of segmentation models.

Feature Engineering: Expanding the feature set for survival prediction models to include additional clinical and imaging features may reduce overfitting and improve predictive accuracy.

Computational Resources: Investment in more advanced computational resources, such as high-end GPUs and increased memory, would facilitate faster and more efficient model training, especially for resource-intensive architectures like 3D Attention U-Net.

10.4 Limitations

The study identified several limitations that require further attention:

- **Missed Segmentation of Tumor Boundaries:** While the 3D U-Net and 3D Attention U-Net models generally performed well in segmenting brain tumors, they occasionally struggled to accurately capture smaller or more complex tumor subregions, such as minor enhancing areas and diffuse edema. Additionally, the models sometimes confused two closely

related classes, as illustrated in **Figure 57** & **Figure 58**. These challenges highlight the need for more advanced feature extraction techniques to better delineate subtle details in medical images.

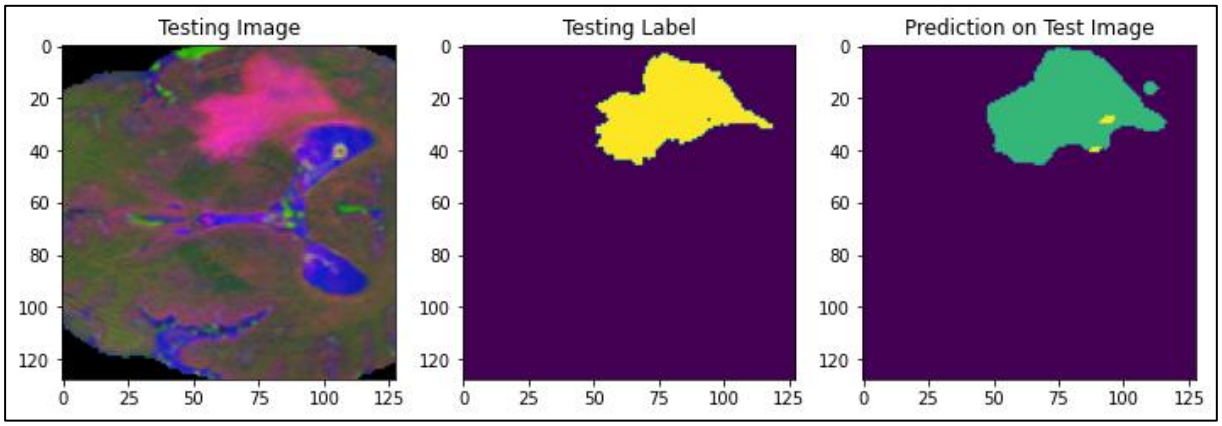


Figure 57. Lack of Accurate Segmentation (3D Attention U-Net)

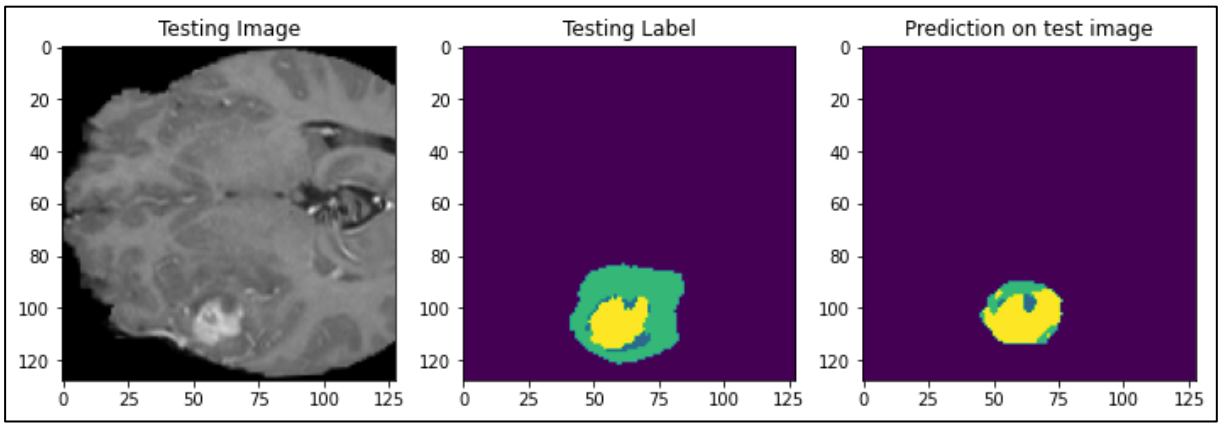


Figure 58. Lack of Accurate Segmentation (3D U-Net)

- Trade-off Between Model Complexity and Computational Resources: The use of complex architectures like 3D Attention U-Net required substantial computational power, limiting the exploration of deeper networks that might have improved segmentation results. Resource constraints also prevented the use of larger batch sizes, potentially affecting model stability and accuracy.

- Feature Limitation in Survival Prediction Models: The survival prediction models were limited by a simplistic dataset that included only basic clinical features, leading to overfitting and reduced generalizability. A more comprehensive feature set, including additional clinical, genetic, or radiomic data, could have enhanced model performance.

- Challenges with Real-Time Application: The high computational and memory demands of the 3D convolutional models restricted their application in real-time clinical settings. More efficient versions, using methods like knowledge distillation, pruning, or advanced transformer architectures, could be developed for environments with limited computational resources.

- Data Augmentation and Generalization Issues: The augmentation techniques used (random flipping, rotation, and zooming) may not have been sufficient to generalize the models across diverse datasets. More advanced strategies, such as GAN-based synthetic data generation or domain adaptation techniques, could better mimic real-world variations and improve model robustness.

10.5 Areas for Future Work

To address these limitations, future work could focus on integrating more advanced architectures, such as transformers combined with convolutional backbones, to enhance feature extraction while maintaining computational efficiency. Additionally, implementing uncertainty estimation techniques could help quantify model confidence in predictions, particularly for critical applications like tumor segmentation. Expanding the data augmentation approach to include novel techniques, such as elastic deformations or more sophisticated synthetic data generation, could further improve model robustness. Lastly, developing a patient-specific modeling approach by incorporating more diverse and personalized data could lead to more accurate and reliable predictions in clinical practice.

9. REFERENCES

Aarshay. (2022, June 15). *Complete Machine Learning Guide to Parameter Tuning in Gradient Boosting (GBM) in Python*. Retrieved from analyticsvidhya: https://www.analyticsvidhya.com/blog/2016/02/complete-guide-parameter-tuning-gradient-boosting-gbm-python/

Abiwinanda, N. H. (2019). Brain tumor classification using convolutional neural network. *In World congress on Medical Physics and Biomedical Engineering*, 183–189.

Afshar, P. M. (2018). Brain tumor type classification via capsule networks. *Proceedings of the 2018 25th IEEE International Conference on Image Processing (ICIP)*, 3129–3133.

Agravat, R. a. (2016). Brain Tumor Segmentation. *Computer Society of India*, (pp. 31-35).

Agravat, R. a. (2018). Deep Learning for Automated Brain Tumor Segmentation in MRI Images. In *Soft Computing Based Medical Image Analysis* (pp. 183-201).

Akay, M. D. (2021). Deep Learning Classification of Systemic Sclerosis Skin Using the MobileNetV2 Model. *IEEE Open Journal of Engineering in Medicine and Biology*, 1-1. doi:doi: 10.1109/OJEMB.2021.3066097

al, A. S. (2021). Evolution in diagnosis and detection of brain tumor. *Journal of Physics: Conference Series*.

Alzubaidi, L. Z.-D.-S.-A. (2021). Review of deep learning: concepts, CNN architectures, challenges, applications, future directions. *Journal of Big Data*, 53.

Amazon AWS. (n.d.). *Introduction to Artificial Intelligence and Machine Learning*. Retrieved from AWS Community: https://community.aws/content/2drbbXokwrIXivItJ8ZeCk3gT5F/introduction-to-artificial-intelligence-and-machine-learning

Bala, P. (2023, Nov 03). *kdnuggets*. Retrieved from Hyperparameter Tuning: GridSearchCV and RandomizedSearchCV, Explained: https://www.kdnuggets.com/hyperparameter-tuning-gridsearchcv-and-randomizedsearchcv-explained

Bankman, I. (2008). *Handbook of Medical Image Processing and Analysis.* Amsterdam: Elsevier.

Bauer, S. W.-P. (2013). A survey of MRI-based medical image analysis for brain tumor studies. *Physics in Medicine and Biology, 58*(13), 97129.

botpenguin. (n.d.). *Transfer Learning* . Retrieved from https://botpenguin.com/glossary/transfer-learning

Bouteille. (2022, Jan 28). *MobileNet-V2: Summary and Implementation*. Retrieved from HackMD: https://hackmd.io/@machine-learning/ryaDuxe5L

BraTS2020 Dataset (Training + Validation). (2020). Retrieved from Kaggle: https://www.kaggle.com/datasets/awsaf49/brats20-dataset-training-validation

Brital, A. (2021, OCT 23). *Inception V3 CNN Architecture Explained* . Retrieved from Medium: https://medium.com/@AnasBrital98/inception-v3-cnn-architecture-explained-691cfb7bba08#id_token=eyJhbGciOiJSUzI1NiIsImtpZCI6ImE0OTM5MWJmNTJiNThjMWQ1NjAyNTVjMmYyYTA0ZTU5ZTIyYTdiNjUiLCJ0eXAiOiJKV1QifQ.eyJpc3MiOiJodHRwczovL2FjY291bnRzLmdvb2dsZS5jb20iLCJhenAiO

Casamitjana, A. P. (2016). 3D convolutional neural networks for brain tumor segmentation: A comparison of multi-resolution architectures . *International Workshop on BraiInternational Workshop on Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries, ,* 150–161.

Chato, L. a. (2017). Machine Learning and Deep Learning Techniques to Predict Overall Survival of Brain Tumor Patients using MRI Images. *Bioinformatics and Bioengineering (BIBE), 2017 IEEE 17th International Conference on. .* IEEE.

DeAngelis, L. (2001). Brain tumors. *New England journal of medicine*, 114-123.

Devasena, L. a. (2013). Efficient computer aided diagnosis of abnormal parts detection in magnetic resonance images using hybrid abnormality detection algorithm. *Central European Journal of Computer Science, 3*(3), 117–128.

Donges, N. (2024). *What Is Transfer Learning? Exploring the Popular Deep Learning Approach*. Retrieved from builtin: https://builtin.com/data-science/transfer-learning

Ertosun MG, R. D. (2015). Automated grading of gliomas using deep learning in digital pathology images: a modular approach with ensemble of convolutional neural networks. . *Annu Symp Proc AMIA Symp*, 1899–1908.

Gates, E. P. (2019). Glioma Segmentation and a Simple Accurate Model for Overall Survival Prediction. In A. B. Crimi, *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries.* Springer, Cham.

geeksforgeek. (2023, Oct 11). *Voting Classifier*. Retrieved from geeksforgeek: https://www.geeksforgeeks.org/voting-classifier/

geeksforgeeks. (2024, June 07). *VGG-Net Architecture Explained*. Retrieved from geeksforgeeks: https://www.geeksforgeeks.org/vgg-net-architecture-explained/#vgg19-architecture

Goswami, S. &. (2013). Bain tumor detection using unsupervised learning based neural network. *International Conference on Communication Systems and Network Technologies*, 573–577.

Gurucharan, M. (2024, June 22). *Basic CNN Architecture: Explaining 5 Layers of Convolutional Neural Network*. Retrieved from upgrad: https://www.upgrad.com/blog/basic-cnn-architecture/

homola, d. (2018, November). *3D U-Net-A TensorFlow implementation of volumetric segmentation*. Retrieved from https://danielhomola.com/unet

IBM. (2024). *What is machine learning (ML)?* Retrieved from IBM: https://www.ibm.com/topics/machine-learning

Isensee, F. K.-H. (2017). Brain Tumor Segmentation and Radiomics Survival Prediction: Contribution to the BRATS 2017 Challenge. *2017 International MICCAI BraTS Challenge.*

Islam, M. J. (2019). Glioma Prognosis: Segmentation of the Tumor and Survival Prediction Using Shape, Geometric and Clinical Information. In A. B. Crimi, *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries.* Springer, Cham.

Japkowicz, N. a. (2015). Performance Evaluation in Machine Learning. In I. L. El Naqa, *Machine Learning in Radiation Oncology* (pp. 85-114). Cham: Springer.

Jovčevska I, K. N. (2013). Glioma and glioblastoma-how much do we (not) know? *Mol Clin Oncol*, 935–935.

Jude Hemanth, D. V. (2014). Performance improved iteration-free artificial neural networks for abnormal magnetic resonance brain image classification. *Neurocomputing, 130*, 98–107.

Kabir Anaraki, A. A. (2019). Magnetic Resonance imaging-based brain tumor grades classification and grading via convolutional neural networks and genetic algorithms. *Biocybern Biomed Eng, 39*(1), 63-74.

Kao, P. N. (2019). Brain Tumor Segmentation and Tractographic Feature Extraction from Structural MR Images for Overall Survival Prediction. In A. B. Crimi, *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries.* Springer, Cham.

Khan HA, J. W. (2020). Brain tumor classification in MRI image using convolutional neural network. *Mathematical Biosciences and Engineering, 17*(5), 6203–6216.

Kittusamy, K. K. (2021). Terrain identification and land price estimation using deep learning. *AIP Conference Proceedings, 2387*, p. 140030. doi:10.1063/5.0068625

Kleihues, P., Burger, P. C., & Scheithauer, B. W. (1993). The new WHO classification of brain tumours. *Brain Pathology, 3*(3), 255-268.

Koehrsen, W. (2018, Jan 10). *Medium*. Retrieved from Hyperparameter Tuning the Random Forest in Python: https://towardsdatascience.com/hyperparameter-tuning-the-random-forest-in-python-using-scikit-learn-28d2aa77dd74

Kori, A. S. (2019). Ensemble of Fully Convolutional Neural Network for Brain Tumor Segmentation from Magnetic Resonance Images. In A. B. Crimi, *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries.* Springer, Cham.

Kshatri, S. &. (2023). Convolutional Neural Network in Medical Image Analysis: A Review. *Archives of Computational Methods in Engineering*.

Liang, C. B. (2018). DRINet for medical image segmentation. *IEEE Transactions on Medical Imaging, 37*(11), 2453-2462.

Liang, Z.-P. a. (2000). *Principles of Magnetic Resonance Imaging: A Signal Processing Perspective.* Bellingham, WA: SPIE Optical Engineering Press.

Louis DN, P. A.-B. (2016). The 2016 World Health Organization Classification of Tumors of the Central Nervous System. *A summary. Acta Neuropathol*, 803 - 820.

Louis, D. O. (2007). The 2007 WHO Classification of Tumours of the Central Nervous System. *Acta Neuropathol*, 97-109.

Mehrotra R, A. M. (2020). A Transfer learning approach for AI-based classification of brain tumors. *Mach Learn Appl, 2*(9), 1-12.

Menze, B. J.-C. (2015). The Multimodal Brain Tumor Image Segmentation Benchmark (BRATS). *IEEE Transactions on Medical Imaging, 34*(10), 1993-2024. doi:10.1109/TMI.2014.2377694

Naveen. (2022, December 10). *What is a confusion matrix?* Retrieved from nomidl: https://www.nomidl.com/machine-learning/what-is-a-confusion-matrix/

Osman, A. (2017). Automated Brain Tumor Segmentation on Magnetic Resonance Images and Patients Overall Survival Prediction Using Support Vector Machines. *International MICCAI Brainlesion Workshop.* Springer, Cham.

Paul, J. &. (2021). Computer aided diagnosis of brain tumor using novel classification techniques. *Journal of Ambient Intelligence and Humanized Computing.*, 12.

Rehman, A. (2024, Jan 21). *Implement 3D-UNet for Cardiac Volumetric MRI Scans in PyTorch*. Retrieved from medium: https://medium.com/@rehman.aimal/implement-3d-unet-for-cardiac-volumetric-mri-scans-in-pytorch-79f8cca7dc68

Rehman, A. N. (2020). A deep learning-based framework for automatic brain tumors classification using transfer learning. *Circuits, Systems, and Signal Processing, 39*(2), 757–775. doi:10.1007/s00034-019-01246-3

Ren, Z. S. (2018). Ensembles of Multiple Scales, Losses and Models for Brain Tumor Segmentation and Overall Survival Time Prediction Task. *International MICCAI Brainlesion Workshop.* Springer, Cham.

Rosebrock, A. (2016, November 7). *Intersection over Union (IoU) for object detection*. Retrieved from pyimagesearch: https://pyimagesearch.com/2016/11/07/intersection-over-union-iou-for-object-detection/

Sachdeva, J. K. (2013). Segmentation, feature extraction, and multiclass brain tumor classification. *Journal of Digital Imaging, 26*(6), 1141–1150.

Sajjad M, K. S. (2019). Multi-grade brain tumor classification using deep CNN with extensive data augmentation. *J. Comput.Sci.*

Saltz J, G. R. (2018). Spatial organization and molecular correlation of tumor-infiltrating lymphocytes using deep learning on pathology images. *Cell Rep, 23*(1), 181–193.

Shin, H. a. (2018). Brain Tumor Segmentation using 2D U-net. *International MICCAI Brainlesion Workshop.* Springer, Cham.

Tahir, B. I. (2019). Feature enhancement framework for brain tumor segmentation and classification. *Microscopy Research and Technique, 82*, 803–811.

Tezcan, B. (2021, June 09). *Why Using a Dummy Classifier is a Smart Move*. Retrieved from towardsdatascience: https://towardsdatascience.com/why-using-a-dummy-classifier-is-a-smart-move-4a55080e3549

The ASCO foundation. (2022). *Cancer-types, "Brain tumor: statistics,"*. Retrieved May 10, 2024, from https://www. cancer.net/cancer-types/braintumor/statistics.

turing. (2024). *Deep Learning vs Machine Learning: The Ultimate Battle*. Retrieved from Turing: https://www.turing.com/kb/ultimate-battle-between-deep-learning-and-machine-learning

Ullah, N. M. (2022). Diabetic Retinopathy Detection Using Genetic Algorithm-Based CNN Features and Error Correction Output Code SVM Framework Classification Model. *Wireless Communications and Mobile Computing*, 1-13.

Varghese, A. S. (2017). Brain Tumor Segmentation from Multi Modal MR images using Fully Convolutional Neural Network. *BRATS proceedings, MICCAI 2017.*

Vinod, R. (2020, May 1). *A detailed explanation of the Attention U-Net*. Retrieved from towardsdatascience: https://towardsdatascience.com/a-detailed-explanation-of-the-attention-u-net-b371a5590831

Xu, X. K. (2018). Brain Tumor Segmentation and Survival Prediction Based On Extended U-Net Model and XGBoost. *International MICCAI Brainlesion Workshop.* Springer, Cham.

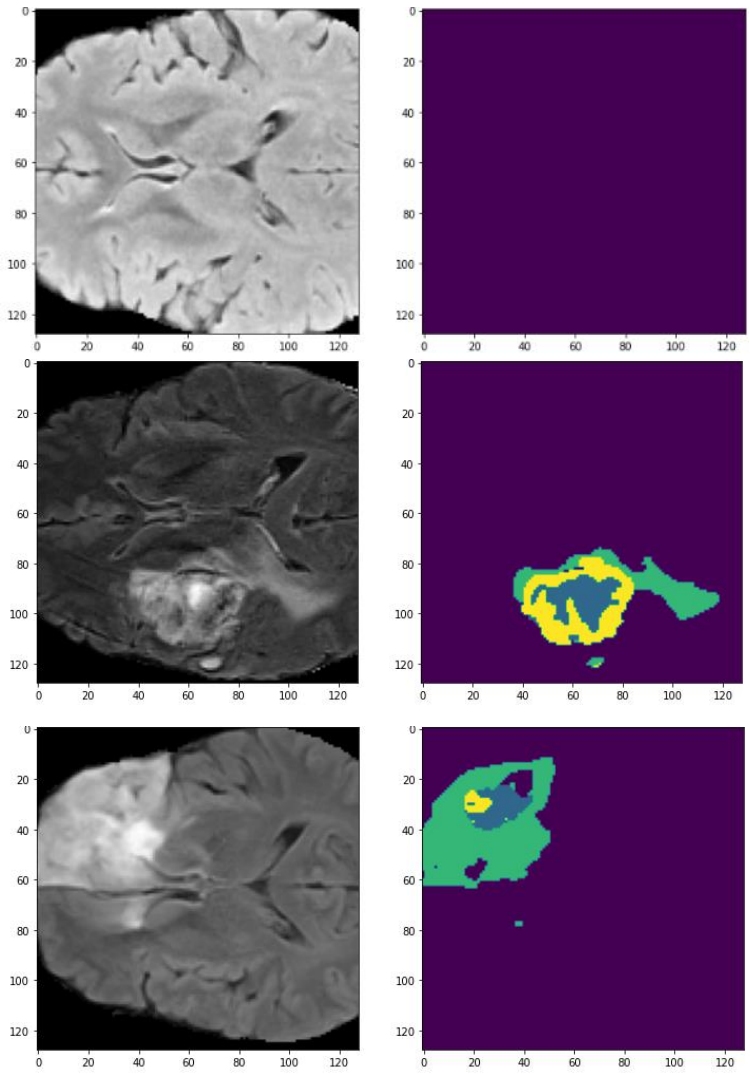
Yang, H. a. (2018). Automatic Brain Tumor Segmentation with Contour Aware Residual Network and Adversarial Training. *International MICCAI Brainlesion Workshop.* Springer, Cham.

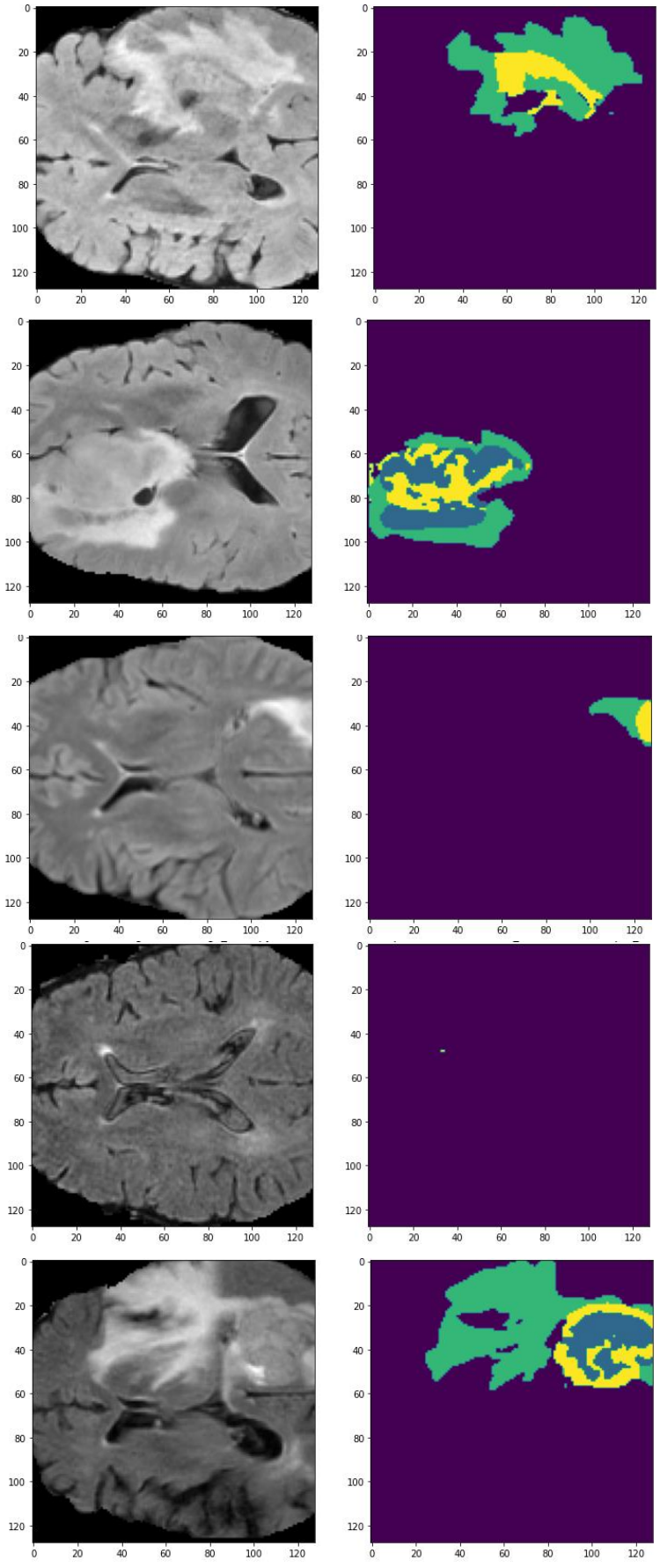
APPENDIX

Results of Brain Tumor Segmentation and Survival Prediction on Unseen Data

This appendix presents the results of brain tumor segmentation and survival prediction performed on unseen MRI data having no target variables (Image mask and survival days). The segmentation results are generated using a 3D U-Net model, which achieved a mean Intersection over Union (IoU) of 75% on the validation dataset, demonstrating its effectiveness in identifying tumor regions. For survival prediction, a Voting Classifier was employed to categorize patients into short, medium, or long survival categories based on the features extracted from the segmented regions and other clinical data.

Segmentation Results :



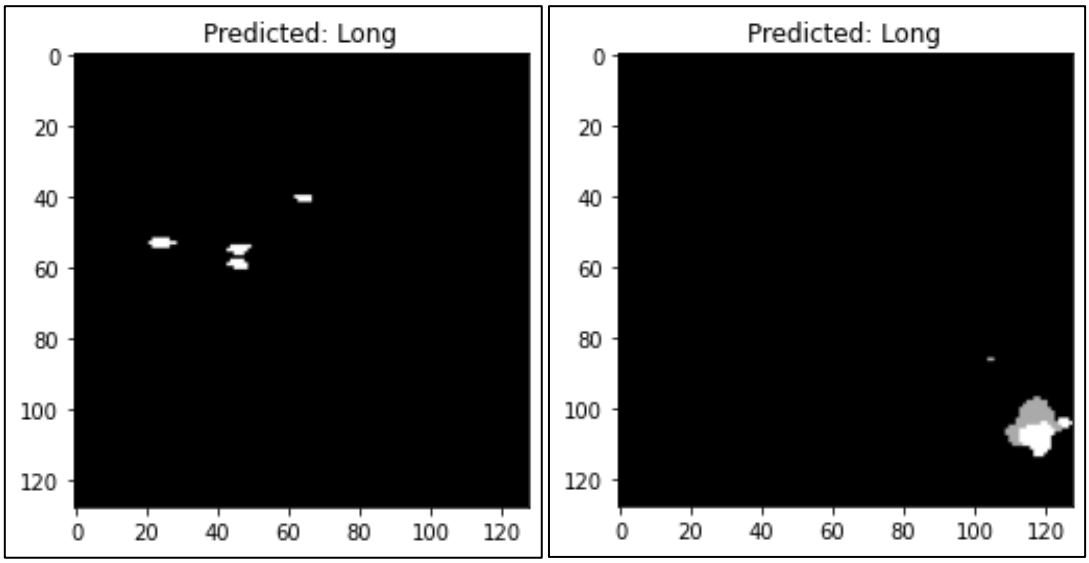


The results above are visualized in pairs of images, with the original MRI slice on the left and the predicted tumor mask on the right. The segmented masks clearly illustrate the regions identified as tumors, demonstrating the model's ability to generalize well to unseen data. The varying shapes and sizes of the segmented regions reflect the model's sensitivity to different tumor characteristics, from small, localized growths to more extensive, diffuse tumors.

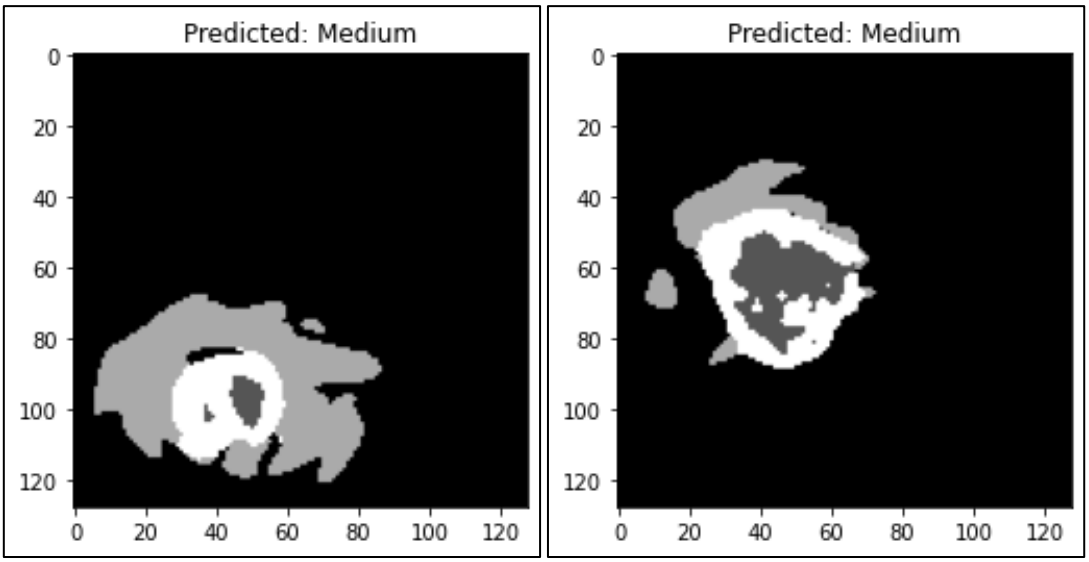
Survival Prediction :

The survival prediction model, based on a Voting Classifier, categorizes patients into Long, Medium, or Short survival groups by analysing MRI features and clinical data. The model shows varying segmentation patterns that influence these predictions:

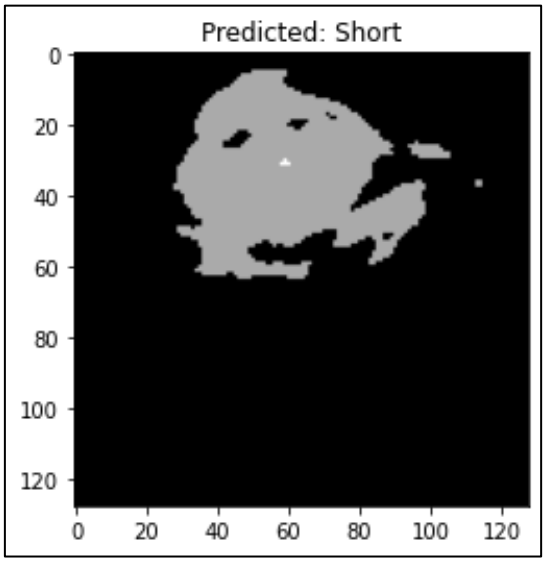
Long Survival: Predictions like those in Figures below are characterized by minimal or sparse segmented regions, indicating less aggressive or smaller tumors in non-critical areas. These cases are associated with a more favorable prognosis and longer survival times.



Medium Survival: Figures below show moderate-sized segmented regions, suggesting a significant but manageable tumor burden. The model predicts medium survival for these cases, reflecting moderate risk and potential for effective treatment.



Short Survival: Predictions such as Figure below feature either no significant segmentation or extensive, widespread regions, indicating aggressive tumor behaviour or critical involvement. These cases correlate with a poor prognosis and shorter survival times.



While the model generally performs well in predicting "Long" survival cases, there is notable overlap in the predicted features for "Medium" and "Short" categories, as well as between "Medium" and "Long" categories. This difficulty in distinguishing between these groups suggests the model struggles when tumor characteristics fall in a borderline range between moderate and severe or moderate and mild involvement. In such cases, the differences in segmented regions may be subtle, or additional non-imaging factors could play a more

significant role in determining the prognosis. These confusions highlight the need for further refinement of the Voting Classifier to enhance its ability to differentiate between these risk levels and provide more accurate survival predictions.