Automated Brain Tumour Diagnosis: Segmentation and Classification with MRI

Leon Leonard - 710019712

Abstract

Despite recent advancements in Machine Learning (ML) and Deep Learning (DL) methods for the classification and segmentation of brain tumours using MRI, there remains a gap in research regarding the integration of these methods for the most common types of brain tumours: gliomas, meningiomas, pituitary, and non-tumours. Furthermore, the impact of segmentation on classification results, particularly in this context, has not been fully explored. This project aims to bridge this gap by exploring the combination of classification and segmentation models for brain tumour diagnosis. The primary aim was to build an automated framework that integrates pre-trained models, including GoogLeNet for classification and U-Net for segmentation, to enhance diagnostic accuracy and accelerate clinical decision making in real-world medical settings. This research explored the benefits in the performance of transfer learning, hyperparameter optimisation, and data augmentations, as well as various loss functions, through careful testing and evaluation. Results of individual models indicate strong performance, with an overall classification accuracy of 98.34%, and overall Dice similarity coefficient of 0.81. However, the direct integration of segmentation into the classification pipeline led to a decrease in classification performance, prompting a shift in the pipeline design. Despite this, the project successfully met its objectives, and with further research and additional resources, remains flexible and open to future enhancements.

I certify that all material in this dissertation which is not my own has been identified.

Contents

1 Introduction

1.1 Background

Brain tumours are one of the leading causes of death among men and women globally, with approximately 126,000 new cases diagnosed each year and a mortality rate of 77% [1]. They reduce life expectancy by an average of 27 years – the highest of any cancer [2]. The impact of untreated brain tumours extends beyond death, as survivors often face a significantly reduced quality of life, due to health risks such as seizures, vision or speech problems, and cognitive defects [3].

There are over 130 known types of brain tumours, each with varying behaviour and treatment plans required. Tumours may be benign - non-cancerous, slow-growing, and non-invasive, or malignant – cancerous, aggressive, and capable of spreading to nearby tissue. The three most prevalent types of primary brain tumours are gliomas, meningiomas, and pituitary tumours. Gliomas are the most common type of malignant tumours, accounting for approximately 81% of malignant brain tumours, and 26% of all diagnosed brain tumours [4]. On the other hand, meningiomas and pituitary tumours are among the most common benign tumours, accounting for 38% and 16% of all primary brain tumours respectively [5]. While malignant tumours like gliomas require prompt treatment, benign tumours also present serious risks. Despite their slower growth, they can gradually increase intracranial pressure and impair brain function, leading to potentially life-threatening outcomes [6].

Given the diverse nature of these tumours, misdiagnosis can have devastating consequences for the patient. Determining whether a tumour is present or not, and hence whether medical intervention is necessary, requires careful and thorough evaluation. This extends to accurately determining the type of tumour if a brain tumour exists. For example, misclassifying a benign tumour as malignant could lead to the patient undergoing aggressive treatments such as chemotherapy or radiotherapy, exposing them to severe side effects and emotional distress [7].

Modern diagnostics increasingly depend on advanced imaging technologies to accurately determine the size and location of brain tumours in a non-invasive approach. Among these, Magnetic Resource Imaging (MRI) has become the gold standard. Its high spatial resolution and exceptional ability in contrasting soft tissues make it especially effective in detecting malformations, without the use of ionising radiation, required by other scans like Computed Tomography (CT) [8]. Furthermore, different imaging modalities can be used in MRI to enhance contrast between different tissues. Despite these advancements, accurate diagnosis still heavily relies on the visual interpretation of MRI scans by radiologists, which is time-consuming, expensive, and prone to subjectivity and human error. Moreover, challenges persist due to the variability in imaging protocols and patient motion artifacts, which can affect the reliability of the results [9].

In response to these challenges, Computer-Aided Diagnosis (CAD) tools have emerged to support tasks such as MRI super-resolution, image segmentation, and image classification, reducing the workload of radiologists and improving diagnostic accuracy [10]. However, CAD systems can be hindered by variability in MRI machine settings, image artifacts, and patient positioning. Recent developments in machine learning (ML) and deep learning (DL) algorithms have shown significant promise in overcoming these limitations.

1.2 Literature review

Early methodologies in brain tumour classification saw the widespread use of traditional ML methods, such as Support Vector Machines (SVM) [11] and k-Nearest Neighbours (KNN). These methods perform well in this context, where a KNN approach [12] achieved an accuracy of 89.5% on The Cancer Imaging

Archive (TCIA), however, ML methods' largest drawback is the reliance on manually defined regions of interest (ROIs) and hand-crafted features. These features, while effective to an extent, struggle to capture the complex patterns present in MRI data, and necessitate expert knowledge for the manual intervention for feature extraction and ROI detection. In the DL field, Convolutional Neural Networks (CNNs) in particular have demonstrated great potential, with their performance optimised through methods such as transfer learning. These models excel in learning hierarchical features from raw MRI images in an unsupervised manner. For the classification of glioma, meningioma and pituitary tumours, research has demonstrated this ability by applying popular CNNs such as AlexNet, GoogLeNet and VGG16, which achieved accuracies of 95.46%, 98.04% and 98.69% respectively [13] on the augmented and contrast-stretched variation of the Figshare dataset [14]. Furthermore, ensemble approaches combining multiple CNN architectures have shown statistically significant improvements in accuracy. One paper aggregated results from AlexNet, VGG16, ResNet18, GoogleNet and ResNet50, to achieve an average accuracy improvement of 2.67% over the individual DL models, and outperformed individual ML methods by 10.12% [15].

Concurrently, DL-based approaches have also revolutionised brain tumour segmentation, where architectures such as CNNs, U-Nets and residual networks have been employed to automatically learn features and isolate tumour regions from MRI scans [16]. For example, one U-Net model trained on the BraTS2021 dataset [17] achieved Dice Similarity Coefficients (DSC) of 90.82%, 82.14% and 77.45% for the whole tumour, tumour core, and enhancing tumour regions respectively [18]. Research in this field has inspired innovations such as the CorrDiff model [19], which integrates a U-Net model with a corrective diffusion model, that helped outperform other state-of-the-art segmentation methods, achieving an average DSC of 86.78%, compared to 80.47% and 84.98% achieved by U-Net and UNETR models respectively on the BraTS2020 dataset. Further research has been conducted into variants of the U-Net model, such as attention U-Nets, which introduce attention gates that show promise in this field [20].

While there is a growing body of research focusing on either the classification or segmentation of brain tumours using deep learning, studies that integrate both tasks into a unified pipeline remain limited, especially on clinically relevant datasets like BraTS2021. Some approaches have attempted to merge these tasks, such as a hybrid deep CNN model for brain tumour classification and detection [21], which segments and classifies types of gliomas. However, these efforts have largely unexplored the classes of glioma, meningioma, pituitary and non tumours. Furthermore, little research has been conducted on the effect of segmentation on the classification of brain tumours, where the isolation of tumours can remove noise and have the potential to improve classification performance. Additionally, the complexity of deep learning models often renders them as "black boxes", which can hinder clinicians' trust in AI systems. A study highlighted that the transparency of AI models in medical imaging varies widely, with scores ranging from 6.4% to 60.9% [22].

This presents a novel and intriguing research question, which is developed and explored in detail in this study.

1.3 Project Aims & Specification

Given the complexity of brain tumour diagnosis and the limitations of traditional methods, this project seeks to develop a transparent, automated pipeline for MRI brain tumour segmentation and classification, to enhance diagnostic accuracy, reduce human error and accelerate clinical decision-making. Such a tool holds the potential to significantly improve patient outcomes by facilitating early and precise diagnosis to enable timely and appropriate therapeutic interventions.

The ultimate aims of this project are as follows:

1. Train a classification model to accurately distinguish the presence of a tumour, and determine the class of tumour with an associated confidence score.
2. Train a segmentation model to accurately identify the tumour region (if present) within an MRI scan.
3. Enhance model accuracy by exploring techniques such as transfer learning, hyperparameter optimisation, and data augmentation, as well as the assessment of different loss functions and evaluation metrics.
4. Investigate the impact of segmentation on classification performance by comparing the accuracy of classification with and without prior segmentation.
5. Integrate segmentation and classification into a fully automated pipeline for robust and efficient brain tumour diagnosis.

2 Design, Methods, and Implementation

2.1 Datasets and Initial Preprocessing

2.1.1 Classification

For the classification task, a comprehensive brain tumour MRI dataset was sourced from Kaggle [23]. This dataset aggregates images from three publicly available datasets: Figshare, SARTAJ, and BR35H into a unified resource, capturing a diverse representation of tumour types and non-tumour cases, and is the most well-populated dataset of its type. Figshare and SARTAJ include labelled instances of glioma, meningioma and pituitary tumours, while BR35H contributes no-tumour cases. The inclusion of no-tumour cases is crucial for diagnostic application, as it enables the model to distinguish between unhealthy and healthy cases, which is vital for avoiding false positives and the psychological distress associated with them.

The dataset has already been split into training and testing sets, with the training set comprising of 1321 glioma, 1339 meningioma, 1457 pituitary, and 1595 no-tumour cases, resulting in a total of 5712 images. The testing set contains 300 glioma, 306 meningioma, 300 pituitary, and 405 no-tumour images, with 1311 in total. Although the classes are imbalanced, there is still a sufficient representation of each and the number of samples per class do not deviate by a large margin. Each image is a 2D MRI slice, captured from either the axial, sagittal, or coronal anatomical planes, and are primarily of the T1-weight contrast-enhanced (T1-CE) modality. A copy of the testing set was created for input into the pipeline, ensuring the pipeline data remains unseen by both the classification and segmentation models.

During initial preprocessing, duplicate entries were detected in both the training and testing datasets. These were removed to prevent data leakage, which could become problematic by introducing bias into the evaluation of the trained model. Image sizes varied from (150, 198) to (1920, 1080), and hence, all images were resized to (256, 256) to standardise input dimensions and align with the downstream model architecture. To evaluate the model's ability to generalise to unseen data during training, 20% of the training set was set aside for validation. This proportion preserves a large amount of training images (4416) to ensure the model learns well, whilst providing enough images (1105) to provide a reliable estimate on the model's predictive performance on unfamiliar data. This data split was performed to ensure similar coverage of each class, avoiding any artificially inflated performance from the over-representation of one class.

2.1.2 Segmentation

For the segmentation task, the BraTS2021 dataset was selected due to its high-quality 3D MRI brain scans, and expert-annotated labelled tumour regions [17]. It's the gold standard benchmarking dataset in brain tumour segmentation, and its widespread application in medical imaging research has encouraged innovation in the exploration of tumour segmentation in this space.

Each subject is represented by four different MRI modalities: T1-weighted (T1), T1-weighted Contrast-Enhanced (T1-CE), T2-weighted (T2), and FLAIR. In addition, each sample includes a ground truth segmentation mask, which categorises tumour regions into four distinct classes - 0: Background, 1: Necrotic and non-enhancing tumour core, 2: Peritumoral edema, 4: Enhancing tumour core. The scope of this project is only concerned with the segmentation of the whole tumour region, and the following classification. Using all four classes may hinder the prediction of each, and therefore the problem is simplified to include two labels: 0 for non-tumour background, and 1 for all tumour regions.

Each MRI modality highlights certain anatomical and pathological features of the brain, offering a holistic view that no single modality could provide alone. However, in order to ensure consistency across the segmentation and classification pipelines, only the T1-CE modality was selected for this study. T1-CE scans show high contrast between active tumour tissue and surrounding structures, making them particularly effective for highlighting growing tumours. This feature should offer greater contextualised information to help distinguish between tumour classes.

The dataset consists of 3108 3D scans, each paired with associated segmentation masks. The data was split into training, validation, and testing sets with a ratio of 70%, 15%, and 15% respectively, resulting in 2175 images for training, 466 for validation, and 467 for testing.

In later experiments, the Figshare labelled segmentation dataset [24] was utilised to expand training data and introduce more relevant brain MRIs to enhance the pipeline. This dataset contains 3064 2D MRI slices with associated binary tumour masks. As Figshare is a subset of the classification dataset, an image hash function was utilised to detect any duplicates, of which 138 were detected. These were removed from the pipeline dataset to maintain the integrity of both sources, and preserve the full set of Figshare images for further training. A 70%/15%/15% split was then applied, resulting in 2144 training, 459 validation, and 461 testing images.

2.2 Methods

2.2.1 Classification

The first classification model explored was GoogLeNet (Inception v1), a 22-layer deep convolutional neural network (CNN) architecture developed by Google researchers [25]. This architecture won the 2014 ImageNet Large Scale Visual Recognition Challenge (ILSVRC), marking a significant advancement in deep learning for computer vision tasks [26]. GoogLeNet is pre-trained on the ImageNet dataset, which consists of a diverse range of natural images, such as animals, vehicles, people, and objects [27].

The core innovation of GoogLeNet is its Inception modules (shown in **Fig. 1**), which apply 1x1, 3x3, and 5x5 convolutions in parallel within the same layer [25]. The filters effectively extract features at different scales and their results are concatenated at each layer. Despite the network being 22 layers deep, GoogLeNet is computationally efficient, due to the integration of 1x1 convolutions and global average pooling, which reduce the number of parameters and effectively replace fully connected layers. Another innovation of GoogLeNet is the use of auxiliary classifiers at intermediate layers of its architecture, directly addressing the vanishing gradient problem often present in training DL models [28]. These characteristics

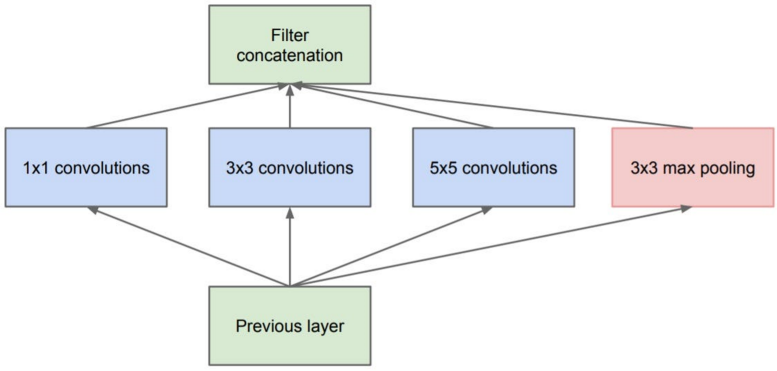


Figure 1: Depiction of Inception v1 module in GoogLeNet.

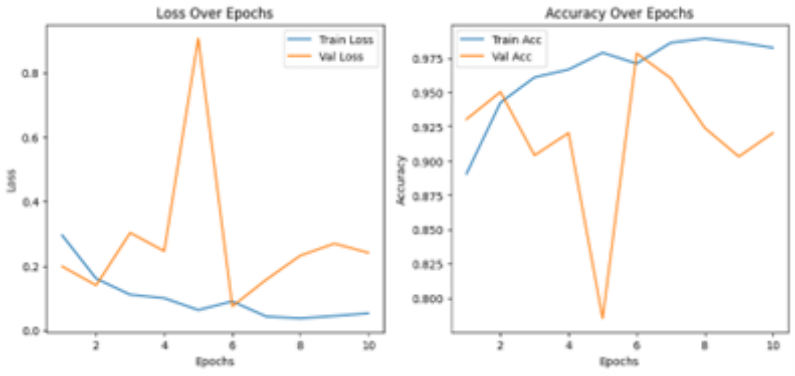
have helped GoogLeNet out-compete other CNN architectures, such as AlexNet, in a variety of different classification problems. Although VGG16 has demonstrated superior performance in the literature, it is significantly more computationally expensive due to its approximately 138 million parameters [29], and was therefore not considered due to the limited resources available for this project.

The pre-trained GoogLeNet model was loaded directly through PyTorch and adapted for this application by adjusting the number of nodes in the output layer to four, corresponding to the four brain tumour types. The input images were normalised to match the distribution the ImageNet dataset (mean: [0.485, 0.456, 0.406], standard deviation: [0.229, 0.224, 0.225]), to ensure consistency and stabilise training. When training the model, the validation accuracy was selected as the primary metric in evaluating the model's performance on generalising to unseen data. The cross-entropy loss function was used due to its effectiveness and efficiency in multi-classification problems. It measures the dissimilarity between the predicted and actual probability distributions, and penalises incorrect predictions more heavily [30].

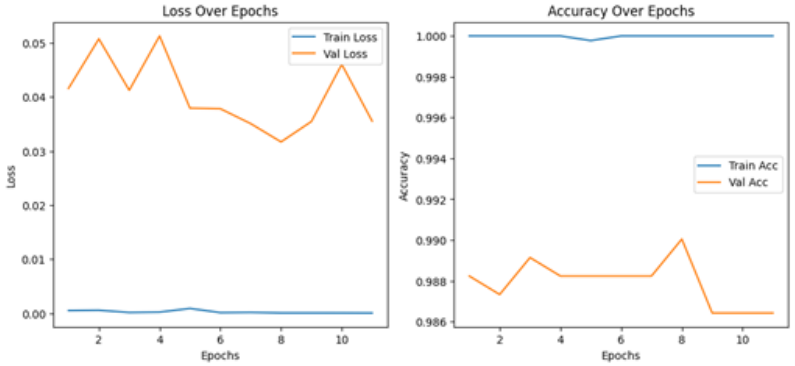
Initially, the model was trained with transfer learning over 10 epochs, with an arbitrary learning rate of 10^{-4} and weight decay of 10^{-8}. The model was optimised with the Adam optimiser, selected for its robust performance and training stabilisation [31]. The model's loss and accuracy on the training and validation sets were tracked for each epoch to monitor convergence. As seen in **Fig. 2**, the model initially learns well up epoch 5, where there is a noticeably sharp decrease in validation accuracy, which stabilises soon afterwards. This suggests instability in learning, likely influenced by the chosen learning rate. The validation accuracy peaks at epoch 6, before steadily decreasing again beyond this, raising concerns of potential overfitting.

Although both training and validation accuracy show strong performance, with peaks of 98.5% and 97.5% respectively, the erratic learning behaviour and early peaking prompted the investigation into hyperparameter tuning and training regularisation strategies to improve stability and generalisation.

To address these concerns, hyperparameter optimisation (HPO) and early stopping were applied. HPO searches for the best combination of hyperparameter values (learning rate, batch size and weight decay), which each directly influence training behaviour. Learning rate controls the update degree of weights, batch size determines how many samples the model processes before updating weights [32], and weight decay is a form of regularisation that controls how much large weights are penalised during training [33]. HPO was run with 20 trials of 50 epochs each, and tested the learning rate, batch size and weight decay in the ranges of $[10^{-5}, 10^{-2}]$, [16, 32, 64] and $[10^{-6}, 10^{-3}]$ respectively. To prevent overfitting, an early



(a) GoogLeNet training behaviour



(b) GoogLeNet training behaviour after HPO

Figure 2: Comparison of Training Performances before and after using Hyperparameter Optimisation, with Cross-Entropy Loss plotted (right) and accuracy plotted (left) for training and validation.

stopper was implemented with a patience of three, which halts training when validation accuracy fails to improve after three epochs. The best set of hyperparameters will be chosen based off the trial that achieves the best validation accuracy.

The optimal set of hyperparameters discovered were: learning rate = 4.33×10^{-5}, batch size = 32, weight decay = 5.95×10^{-6}. Training with these methods yielded more stable training and convergence, with higher evaluation performance as captured in **Fig. 2**. Training loss began at 0.001 and remained very low with only slight fluctuations. This may suggest that the optimised learning rate applied useful pre-trained knowledge well that is relevant to brain tumour features. Validation loss decreased from just above 0.04 and to around 0.03, where it plateaued and early stopping was triggered at epoch 11. This corresponded to a peak validation accuracy of 0.99, demonstrating the effectiveness of HPO and early stopping in generalisation.

2.2.2 Segmentation

The chosen segmentation model is a pre-trained U-Net trained on brain MRI scans [34]. This model was specifically trained on FLAIR sequences from the TCIA LGG collection and achieved a strong median DSC of 87.33%. Although the model was trained on FLAIR sequences, the model architecture and learned tumour features should transfer well to T1-CE sequences due to the shared tumour characteristics between these modalities. Leveraging this pre-trained model greatly improves computational efficiency, whilst still retaining high-quality feature representations from the same domain.

To ensure compatibility with the model, the BraTS2021 images were converted from 3D volumes of

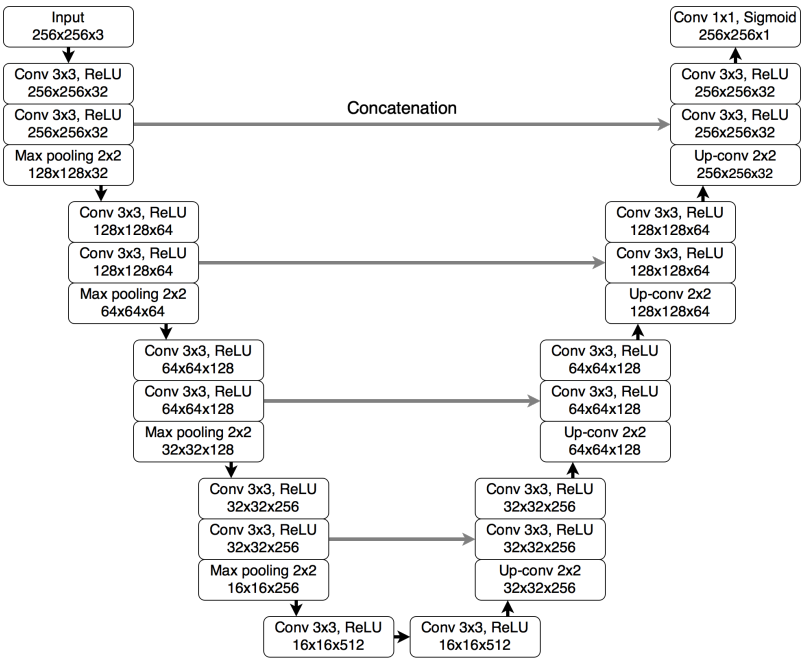


Figure 3: Architecture of the standard U-Net model for image segmentation.

shape (240x240x155) to 2D slices, and resized to shape (256x256). Although MRI scans are inherently 3D because they capture a series of 2D slices stacked together, using a single 2D slice is far easier to work with. One slice was taken from each anatomical plane: axial, coronal, and sagittal, which correspond directly to the images in the classification dataset, ensuring cross-model consistency. The SciPy center_of_mass function was applied to the tumour mask tensor to identify the 'best' slice for each plane, or where the tumour is at its greatest representation. Each slice was duplicated across all three RGB channels, the same approach taken by the authors of this model. In no-tumour cases, the central slice across each plane was captured instead. The final pre-processing step involved normalising the slices with a minmax function to ensure consistency.

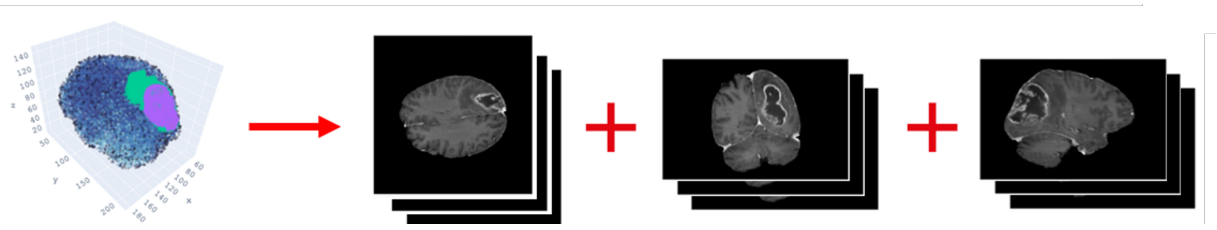


Figure 4: Depiction of 3D volumes converted into 2D slices.

For segmentation, the task is framed as a pixel classification problem, whereby individual pixels are labelled as either a tumour (1), or non-tumour region (0). The model outputs raw probabilities, which are thresholded at 0.5 to convert them into binary class predictions. During training, however, the raw probabilities are retained to preserve contour information for loss calculation.

Three different loss functions were evaluated:

- BCELoss: Measures the difference between the true labels (0 or 1) and the predicted probabilities, commonly used for binary classification tasks.
- DiceLoss: Focused on maximising spatial overlap between predicted and ground truth tumour

regions.

- CombinedLoss (BCELoss/DiceLoss + BoundaryLoss): This custom loss function combines the BCE/Dice Loss with Boundary Loss to enhance the model's sensitivity in capturing the contours and edges of the tumour region. This implementation of BoundaryLoss uses 3x3 Sobel filters, to detect edges in horizontal and vertical directions.

The primary evaluation metric was selected as the Dice similarity coefficient (DSC), with the 95th percentile of Hausdorff distance (HD95) as a complementary metric. The DSC measures the overlap between the predicted and ground truth regions, whereas the Hausdorff distance measures the greatest Euclidean distance from the predicted segmentation to the closest point in the ground truth segmentation. A combined metric was computed with a ratio DSC/HD95 of 0.7/0.3, to complement each other on the overlap and edge detection of predictions.

2.3 Experimental Design

For classification, each model will be evaluated on the testing subset of the classification dataset, based on their accuracy, precision, recall, and F-1 scores. The fine-tuned models will be benchmarked against the baseline model across these metrics. The classification experiments will compare the effects of transfer learning and hyperparameter optimisation on the prediction of unseen data.

For the segmentation experiments, model performance will be assessed on the BraTS2021 testing dataset, focusing on their overall DSCs as well as their ability to handle challenging images. For experiments with additional slices, evaluation will be performed on only centre slices to ensure consistency and fairness across models. The combinations of the loss functions with boundary loss will be of weighting 0.7/0.3, to prioritise overall pixel classification while detecting tumour contours.

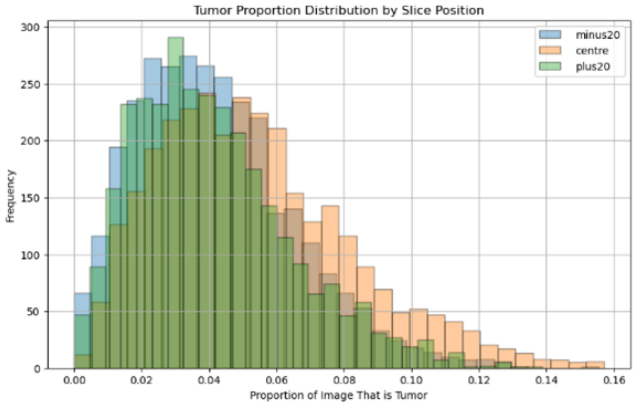
The first segmentation experiment involved training the pre-trained U-Net model on the BraTS2021 slices, and comparing its performance to the baseline. This experiment was designed to explore how transfer learning and hyperparameter optimisation would affect segmentation performance. For this study and beyond, each model was trained with and without hyperparameter optimisation, and their counterparts were compared. Each model was initially trained with a learning rate of 10^{-4}, batch size of 32 and weight decay of 10^{-5}, with the same hyperparameter search space as classification. The Adam optimiser was employed, and an early stopper was also implemented with a patience of 3 on the validation DSC. Each HPO experiment was conducted with 20 trials (reduced to 10 for experiments involving multiple slices due to computational constraints), with 50 epochs per trial.

The next set of experiments compared the effects of different loss functions on segmentation quality. The best-performing model from this phase was carried forward to conduct further experiments. The following experiment investigated whether the use of an augmented dataset would improve the generalisability of the model. Two approaches for augmentations were considered: applying preprocessing techniques such as biased crop, Guassian blur and rotation, or taking multiple slices in each plane to capture the tumour where its shape may not as well defined. The latter approach was selected due to its closer alignment with real-world MRI practices, where tumour slices are often manually selected, and cannot guarantee the most representative slice will be captured. Three slices were extracted from each plane: the centre_of_mass slice, a slice taken 20% anterior to the centre, and a slice taken 20% posterior to the centre. This was applied to all data splits and resulted in a total of nine slices from each subject.

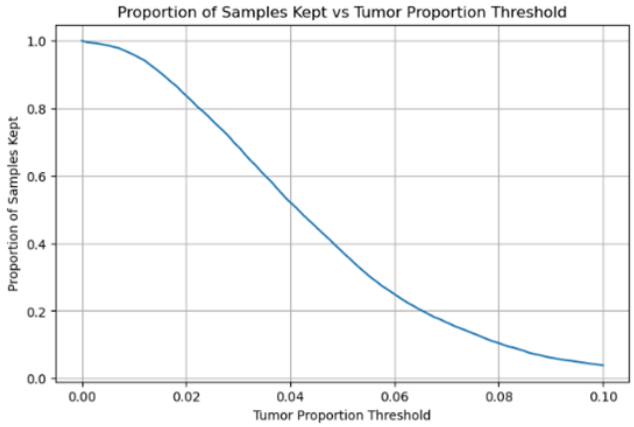
A further investigation was conducted to assess whether extremely small tumour regions negatively bias the model's training performance. **Fig. 5a** depicts the tumour proportion distributions for minus20,

centre and plus20 slices. As expected, the centre slices have on average the greatest coverage. However in edge cases (particularly with minus20 and plus20 slices), tumours were either minimally visible or entirely absent, as illustrate in **Fig. 6**. These extreme cases could hinder training if too much emphasis is placed on them, as it could introduce noise and cause the model to overfit to less informative regions. To mitigate this, images with minimal coverage were removed by applying two thresholds: 0.01 and 0.005 tumour coverage.

Applying a threshold of 0.005 resulted in 9197 total images, whereas applying a stricter threshold of 0.01 resulted in 8937 total images. In both cases, the resulting datasets remained sufficiently large enough for robust training and evaluation (as seen in **Fig. 5b**, with 98% and 95% of the original dataset for 0.005 and 0.01 thresholds respectively. This thresholding step contributes to model stability, and the results of both will be compared.



(a) Histogram of Tumour Coverage Proportion for 20% anterior, centre, and 20% posterior slices



(b) Proportion of Samples Kept by Tumour Proportion Threshold

Figure 5: Tumour coverage statistics visualised in histogram and threshold retention curves.

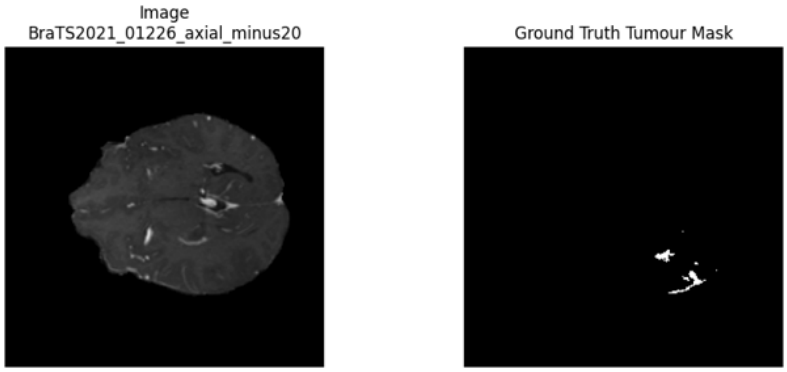


Figure 6: Example of Small Tumour Coverage.

Once the best-performing segmentation model was selected, it was integrated with its saved weights into the final pipeline. The initial investigation focused on assessing the impact of segmentation on classification accuracy.

In this experiment, the segmentation model was incorporated into the classification pipeline. All images in the classification dataset were preprocessed in-line with the segmentation data before being

passed through the model. The segmented tumour images were then further preprocessed in-line with the classification pipeline and were used to train the GoogLeNet model. No-tumour images were also processed through the segmentation model to avoid bias. The model's performance was evaluated on classification accuracy and confidence scores, and compared against the best-performing GoogLeNet model fine-tuned directly on raw data.

2.4 Pipeline Design

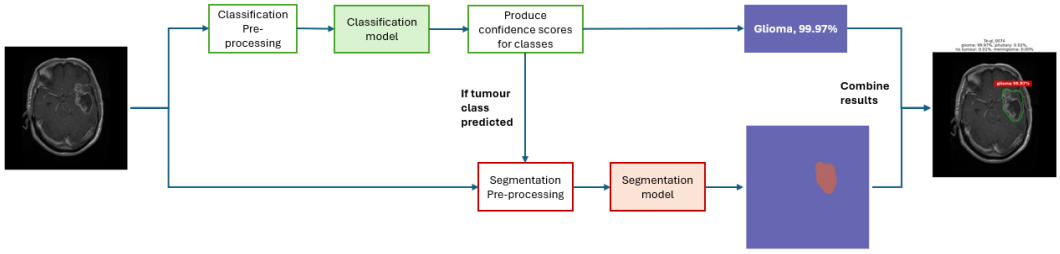


Figure 7: Flow Diagram of the Pipeline Design.

The pipeline data is the testing set of the MRI classification dataset, with duplicate entries shared with Figshare removed, comprising of 1173 images. The pre-trained classification and segmentation models are first loaded in evaluation mode. Each image is then pre-processed for classification and segmentation with their respective pre-processing steps, and the classifier makes its prediction with an associated confidence score. If the predicted class is not "no-tumour", the tumour mask will be generated through the segmentation model. The final output displays the MRI scan, with an outlined tumour region (if present), and its predicted class and confidence score labelled above the area. For transparency, the confidence scores of all classes will be provided, as they are crucial in real-world medical applications. By revealing the model's uncertainty, these scores will allow clinicians to assess the reliability of predictions and make informed decisions on patient care.

2.5 Implementation

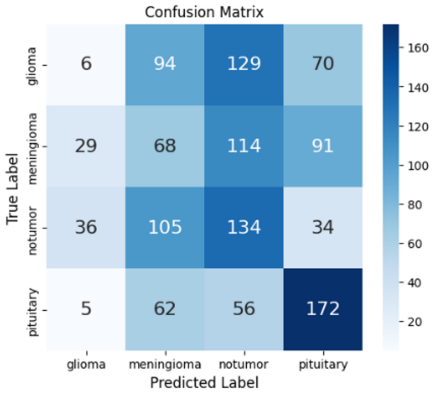
This project was developed using Python, utilising the PyTorch framework for training and evaluating deep-learning models. Key libraries were used such as NumPy for data manipulation, Matplotlib for producing visualisations, and Optuna for HPO. Training was performed on virtual machines, including Google Colab and Azure VMs, utilising the T4 GPU to accelerate performance. Throughout these experiments, several strategies were used to optimise data loading and processing, including storing images in compressed NumPy formats, using memory-mapped loading, and lazy data loaders. These optimisations improved training speed, enhanced memory efficiency, and enabled more experiments to be conducted with the given time and resource constraints.

3 Results and Analysis

The following section presents the results of the individual classification and segmentation models, the effect of segmentation on classification, and the final integrated pipeline.

3.1 Classification

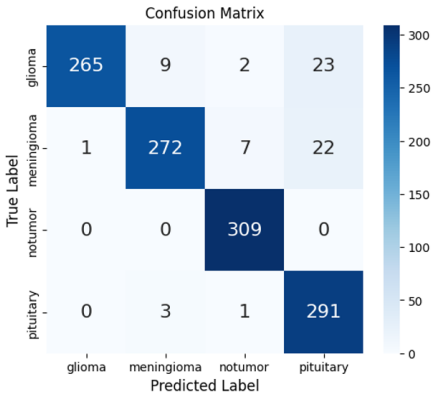
The first experiment focused on evaluating the classification model's ability to differentiate between various brain tumour types, by leveraging the GoogLeNet architecture fine-tuned on the MRI dataset. All results can be seen in **Fig. 11**, with the confusion matrices and classification reports for each model



(a) Confusion Matrix for Baseline GoogLeNet

	precision	recall	f1-score	support
glioma	0.08	0.02	0.03	299
meningioma	0.21	0.23	0.22	302
notumor	0.31	0.43	0.36	309
pituitary	0.47	0.58	0.52	295
accuracy			0.32	1205
macro avg	0.27	0.32	0.28	1205
weighted avg	0.27	0.32	0.28	1205

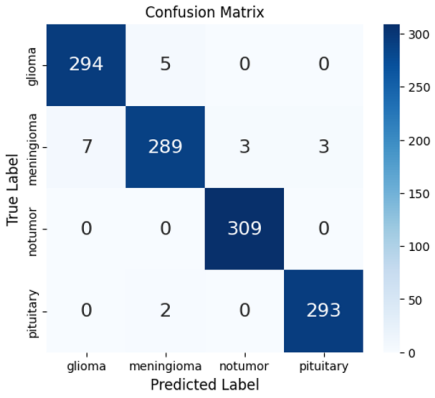
(b) Classification Report for Baseline GoogLeNet



(c) Confusion Matrix for Fine-tuned GoogLeNet

	precision	recall	f1-score	support
glioma	1.00	0.89	0.94	299
meningioma	0.96	0.90	0.93	302
notumor	0.97	1.00	0.98	309
pituitary	0.87	0.99	0.92	295
accuracy			0.94	1205
macro avg	0.95	0.94	0.94	1205
weighted avg	0.95	0.94	0.94	1205

(d) Classification Report for Fine-tuned GoogLeNet



(e) Confusion Matrix for Fine-tuned GoogLeNet with HPO

	precision	recall	f1-score	support
glioma	0.98	0.98	0.98	299
meningioma	0.98	0.96	0.97	302
notumor	0.99	1.00	1.00	309
pituitary	0.99	0.99	0.99	295
accuracy			0.98	1205
macro avg	0.98	0.98	0.98	1205
weighted avg	0.98	0.98	0.98	1205

(f) Classification Report for Fine-tuned GoogLeNet with HPO

Figure 8: Comparative performance of GoogLeNet models for brain tumor classification. (a) and (b) show results for the baseline model, (c) and (d) for the fine-tuned model, and (e) and (f) for the model with hyperparameter optimization.

variant. The baseline model (pre-trained on ImageNet) performed poorly, achieving an overall accuracy of 32% with significant missclassifications. This is particularly prevalent for glioma, with a recall of just 2%. The confusion matrix further illustrates the imbalance in class predictions, with a large proportion of glioma cases misclassified as no-tumour. However, this was to be expected, as the baseline model had not been fine-tuned for the specific task of brain tumour classification.

After fine-tuning the model on the brain MRI dataset, the classification accuracy improved substantially, achieving an overall accuracy of 94%. Precision, recall, and F1-scores across all tumour types showed extremely large improvements, with glioma cases showing the most notable gain, achieving a precision and recall of 1.0 and 0.89 respectively, and an F1-score of 0.94. The model's lower recall relative to precision suggests that it is quite conservative with its glioma predictions, with 23 out of 34 missed cases being misclassified as pituitary. In contrast, for no-tumour cases, the model achieved a perfect recall of 1, correctly identifying all no-tumour instances, which carries immense importance into a clinical setting as it avoids the overdiagnosis of patients. These results highlight the strength of transfer learning, where the model has adapted its learned features to the specific characteristics of tumours in brain MRI scans.

After training with HPO, the model achieved further gains over its fine-tuned counterpart, with an exceptional overall accuracy of 98.34%. The alignment between precision and recall was much closer for all classes. Although its precision dropped slightly to 0.99, the recall of glioma cases rose to 0.97. Furthermore, all non-tumour cases were still correctly identified. The improved performance with HPO indicates that the optimally chosen learning rate, batch size, and weight decay, prevented the model from overfitting while maintaining high accuracy on both the training and testing datasets.

While the classification results demonstrate significant improvements, there still exist some prediction errors which leaves room for further research and development. Initially, the exploration of additional classifiers or applying data augmentations were considered, but it was decided that prioritising the development of the segmentation algorithm and the integration of the full pipeline would yield greater benefits towards the ultimate goals of this project.

3.2 Segmentation

The full set of segmentation experiments and their results can be seen in **Table. 1**. The baseline model performed very poorly on the evaluation data, achieving a median Dice score of 0.0995. Despite being pre-trained on brain MRI images, the differences in training data and MRI modality likely limited the model's ability to generalise well. However, upon fine-tuning the model on the BraTS2021 dataset, it was able to adapt considerably well, with the median Dice score increasing to an average 0.77 across all standalone loss functions.

In **Fig. 9**, the predicted tumour regions of the BCELoss-optimised model are visually compared with and without HPO to the ground truth tumour mask. In the first two cases, dice scores exceeded 0.9, and the overall tumour coverage was strong. The tumour contours were relatively well-captured, but optimising for tumour coverage led to the underrepresentation of smaller tumour regions, as seen in the sagitall image. In contrast, with more stable learning attained in HPO, the model performed slightly worse in terms of Dice score but captured the contours and sub-tumour regions more effectively. While these split-tumour cases are less common (and perhaps their underrepresentation in training can explain the models' struggles), this would still be problematic in clinical applications where accurate boundary detection is critical. Thus, striking a balance between overlap and edge detection is crucial in producing more accurate and versatile segmentation results.

Experiment	HPO	Dataset Variant	Dice Score (Median ± Std)
Baseline	N/A	Centre slices	0.0995 ± 0.0478
BCELoss	False	Centre slices	0.7787 ± 0.2105
BCELoss	True	Centre slices	0.7624 ± 0.2228
DiceLoss	False	Centre slices	0.7707 ± 0.2478
DiceLoss	True	Centre slices	0.7559 ± 0.2334
BCELoss + BoundaryLoss	False	Centre slices	0.8057 ± 0.1958
BCELoss + BoundaryLoss	True	Centre slices	0.8097 ± 0.1854
DiceLoss + BoundaryLoss	False	Centre slices	0.7441 ± 0.2630
DiceLoss + BoundaryLoss	True	Centre slices	0.7641 ± 0.2432
BCELoss + BoundaryLoss	False	Multiple slices	0.7938 ± 0.2103
BCELoss + BoundaryLoss	True	Multiple slices	0.7985 ± 0.2030
DiceLoss + BoundaryLoss	False	Multiple slices	0.7846 ± 0.2331
DiceLoss + BoundaryLoss	True	Multiple slices	0.8126 ± 0.1877
BCELoss + BoundaryLoss	False	Multiple slices (0.005 threshold)	0.7747 ± 0.2240
BCELoss + BoundaryLoss	True	Multiple slices (0.005 threshold)	0.8090 ± 0.1993
BCELoss + BoundaryLoss	False	Multiple slices (0.01 threshold)	0.8106 ± 0.1798
BCELoss + BoundaryLoss	True	Multiple slices (0.01 threshold)	0.8061 ± 0.1886
DiceLoss + BoundaryLoss	False	Multiple slices (0.005 threshold)	0.8057 ± 0.2132
DiceLoss + BoundaryLoss	True	Multiple slices (0.005 threshold)	0.8001 ± 0.2193
DiceLoss + BoundaryLoss	False	Multiple slices (0.01 threshold)	0.8012 ± 0.2117
DiceLoss + BoundaryLoss	True	Multiple slices (0.01 threshold)	0.8107 ± 0.2017

Table 1: Experimental results comparing different loss functions, HPO settings, and dataset variants

This insight motivated the incorporation of the Boundary Loss function, which was coupled with both Dice Loss and BCE loss in subsequent experiments. Although the models trained with these losses showed an improvement in overall Dice score, they continued to struggle with transferability in more difficult cases, such as the sagitall example, where the sub-tumour region remained poorly captured.

To address this challenge and improve the models' generalisability, multiple slices were taken from each plane to diversify the training set with more difficult images, where the tumour region may not be as easily recognisable. As a result, the evaluation Dice scores showed differing performances between the DiceLoss + BoundaryLoss and BCELoss + BoundaryLoss models, with the former showing an improvement, while the latter experienced a slight deterioration in performance. Neverless, both models not only identified the sub-tumour region, but also improved the contours of the other testing samples, reflecting

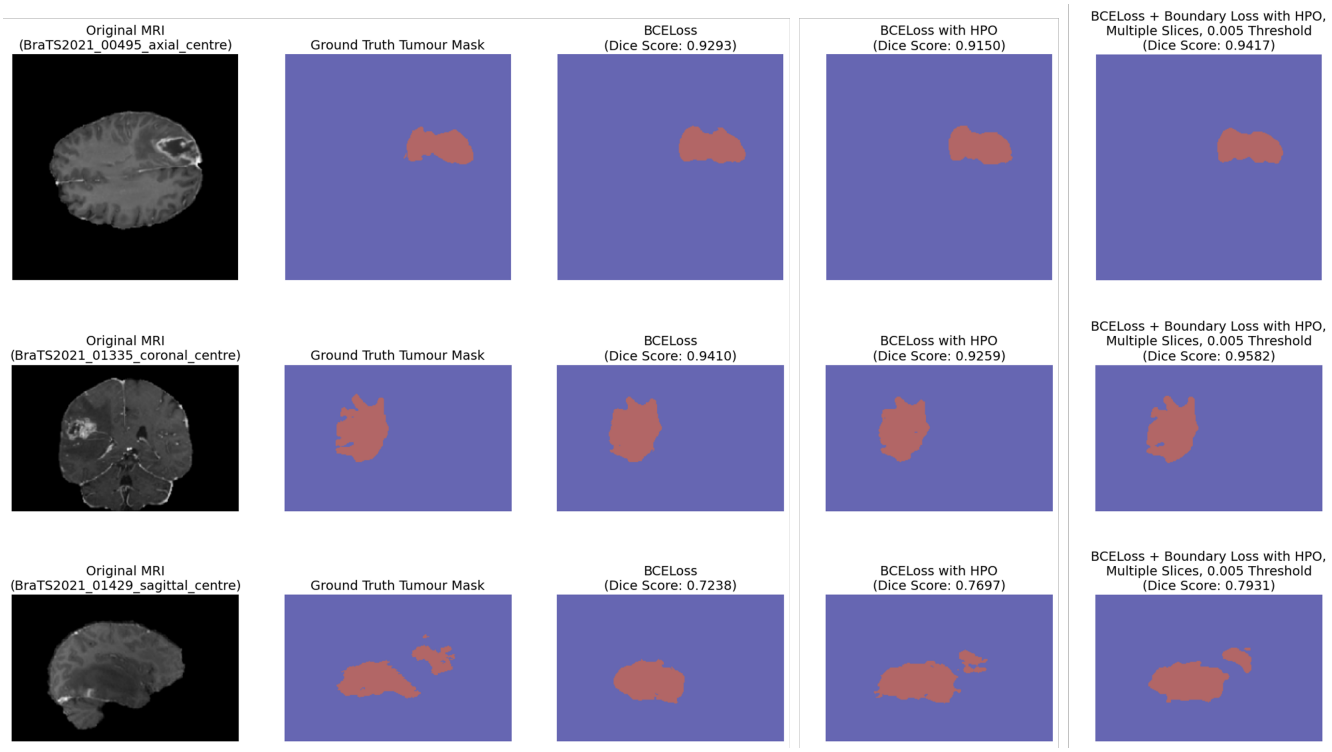


Figure 9: Visual comparison of brain tumor segmentation results across different loss functions, with and without Hyperparameter Optimization (HPO), and with varying dataset variants, shown for axial, coronal, and sagittal views.

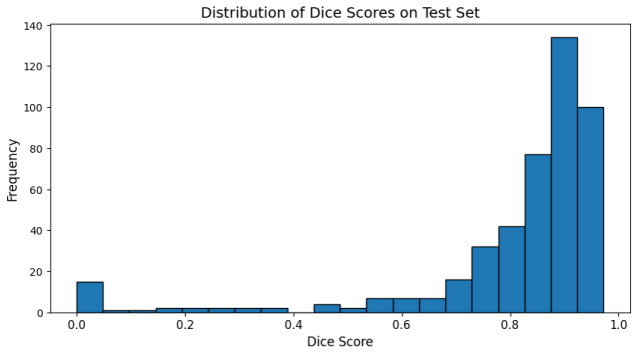
the benefits of including a broader range of training data.

Although the addition of less favourable slices improved model performances, it also introduced a large number of very small or poorly represented tumour regions. If training focuses too heavily on capturing features of these regions, its learned weights may bias towards these anomalous and non-representative cases, thus negatively affecting model learning and performance. To mitigate this, further experiments were conducted where images with a tumour coverage smaller than 0.01 or 0.005 were removed from the dataset. This change led to further improvements for both models, where the enhanced model produces much closer-resembling tumour masks, and more refined contour accuracy, as seen in **Fig. 9**.

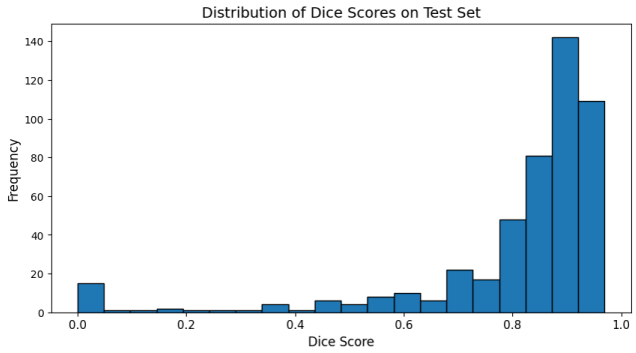
After conducting all experiments with careful evaluation, the BCE loss + Boundary loss HPO model, trained on multiple slices with a 0.005 threshold, was selected for the final pipeline. Although its median Dice score of 0.809 is considerably strong, it's slightly lower than the highest (0.8107), achieved by the Dice loss + Boundary loss HPO model on multiple slices with a 0.01 threshold. However, given the minimal difference in Dice scores, the former model was preferred due to the larger and more diverse training set, and it still exhibited a slightly better dice score distribution, with points skewed left (**Fig. 10**). Despite measures taken in place, there are still some cases where the mask overlap is very minimal, which may become problematic in the final pipeline.

The key insights taken from these experiments include the importance of increasing and diversifying the training set, striking a balance between tumour overlap and edge detection, and removing abnormalities. These strategies collectively contributed to improving model stability and performance. Although the improvement in median Dice score was not dramatic, the model's ability to capture tumour contours and sub-tumour regions improved greatly, without compromising overall performance. The gains

in generalisability allow the model to perform more robustly across various anatomical planes, which is critical for real-world applications, where the model will need to handle diverse and challenging MRI scans.



(a) DiceLoss + BoundaryLoss with HPO, multiple slices, 0.01 removal



(b) BCELoss + BoundaryLoss with HPO, multiple slices, 0.005 removal

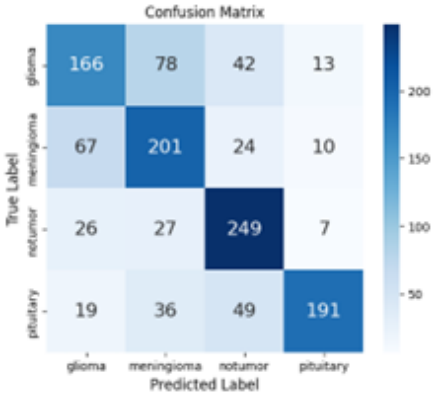
Figure 10: Comparison of dice distribution histograms on centre slices.

3.3 The Effect of Segmentation on Classification

After selecting the best-performing segmentation model, the fine-tuned model was incorporated into the classification pipeline to assess its impact on classification performance. Without HPO, the classification model achieved an overall accuracy of 67%. The most well-predicted class was no-tumour, with a recall of 0.81, whereas the weakest predicted class was glioma, with a recall of 0.56. These results reflect the performance of the original classification model, albeit with significantly lower results. With the introduction of HPO, the overall accuracy slightly dropped to 66%, however its classification of pituitary and glioma cases improved, with F1-scores of 0.77 and 0.59 respectively. The performance of the HPO-enhanced model was still substantially lower than the best-performing GoogLeNet fine-tuned on the raw images.

For the pipeline, the classification model generates confidence scores by applying the softmax function to the raw logits for each class, ensuring that the total probability sums to 1. For further investigation here, the distribution of confidence scores for each model was plotted in **Fig. 12**. The graph reveals that the fine-tuned GoogLeNet on raw images exhibits high confidence in its predictions, with most scores concentrated in the [0.97, 1.0] range. In contrast, the GoogLeNet models fine-tuned on the segmented images show a much broader distribution of confidence scores, suggesting far less predictability.

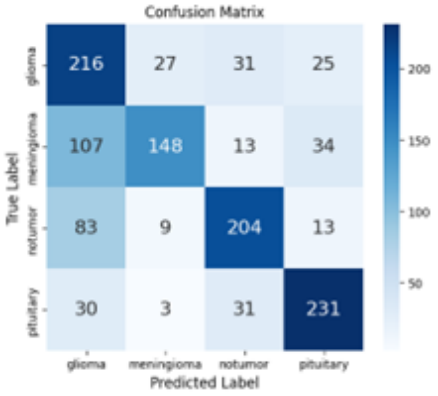
These results provide evidence that the use of segmentation as a preprocessing step may be detrimental to classification accuracy. This can be largely attributed to the loss of spatial information around the tumour, that may contain valuable contextual features. This is because the features of brain tumours are not just determined by size and shape, but also the location the tumour develops from, such as pituitary tumours, which distinctly only grow around the pituitary gland. GoogLeNet is pre-trained on natural images and learn to extract general features, so when fine-tuned, it can effectively process whole brain MRI scans without the need for segmentation, as they leverage their ability to capture broader patterns across the entire image. Additionally, the results produced by the segmentation model are prone to errors, and the aggregation of these reasons explain why the segmented approach underperformed in comparison to using the raw images.



(a) Confusion Matrix for Fine-Tuned GoogLeNet on Segmented Images

	precision	recall	f1-score	support
glioma	0.60	0.56	0.58	299
meningioma	0.59	0.67	0.62	302
notumor	0.68	0.81	0.74	309
pituitary	0.86	0.65	0.74	295
accuracy			0.67	1205
macro avg	0.68	0.67	0.67	1205
weighted avg	0.68	0.67	0.67	1205

(b) Classification Report for Fine-Tuned GoogLeNet on Segmented Images



(c) Confusion Matrix for Fine-Tuned GoogLeNet on Segmented Images with HPO

	precision	recall	f1-score	support
glioma	0.50	0.72	0.59	299
meningioma	0.79	0.49	0.61	302
notumor	0.73	0.66	0.69	309
pituitary	0.76	0.78	0.77	295
accuracy			0.66	1205
macro avg	0.70	0.66	0.66	1205
weighted avg	0.70	0.66	0.66	1205

(d) Classification Report for Fine-Tuned GoogLeNet on Segmented Images with HPO

Figure 11: Comparative performance of GoogLeNet models for brain tumor classification on segmented images. (a) and (b) show results without HPO, and (c) and (d) show results with HPO

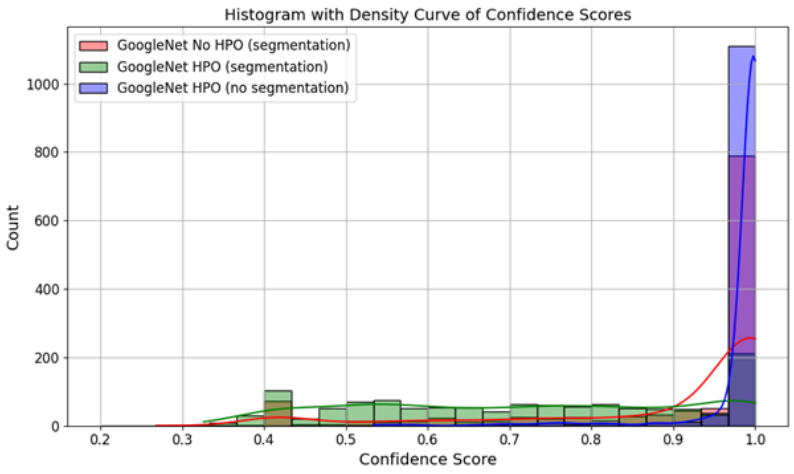


Figure 12: Histogram with density curve showing the distribution of confidence scores across different GoogLeNet model configurations.

3.4 Pipeline Results

The chosen segmentation model was initially applied to the pipeline data, a subset of the MRI classification testing dataset. However, due to the isolated nature of brain structures in the BraTS2021 dataset, the model frequently confused non-tumour anatomical regions, such as eyeballs and other parts of the head, with the tumour itself. This often resulted in over-segmentation or, in some cases, missing the tumour region entirely. These issues are particularly evident in **Fig. 13**, which presents examples of segmentation results where anatomical structures were mistakenly identified as tumour regions.

This prompted the investigation of how to ensure greater cohesion between the two datasets, in order to produce more reliable and applicable results. The first consideration was re-training with a more advanced U-net model, however the root cause appeared to lie in the dataset itself. To improve this, the Figshare segmentation dataset was discovered, and it closely matched the appearance of MRI scans in the classification dataset.

The segmentation model was then further fine-tuned on the Figshare dataset. By incorporating the learned features from BraTS2021 with the more relevant data from Figshare, the model could adapt more effectively to the types of images it would be exposed to in real-world settings, where pre-processing steps like skull-stripping are often not performed prior to analysis. Although the segmentation model could've been trained solely on the Figshare dataset, leveraging BraTS2021 slice augmentations enhanced the models ability to capture diverse tumour shapes and regions. Despite the model being trained on slices with 0.005 removal, this was not conducted for the Figshare dataset to ensure best transferability to data in a similar domain.

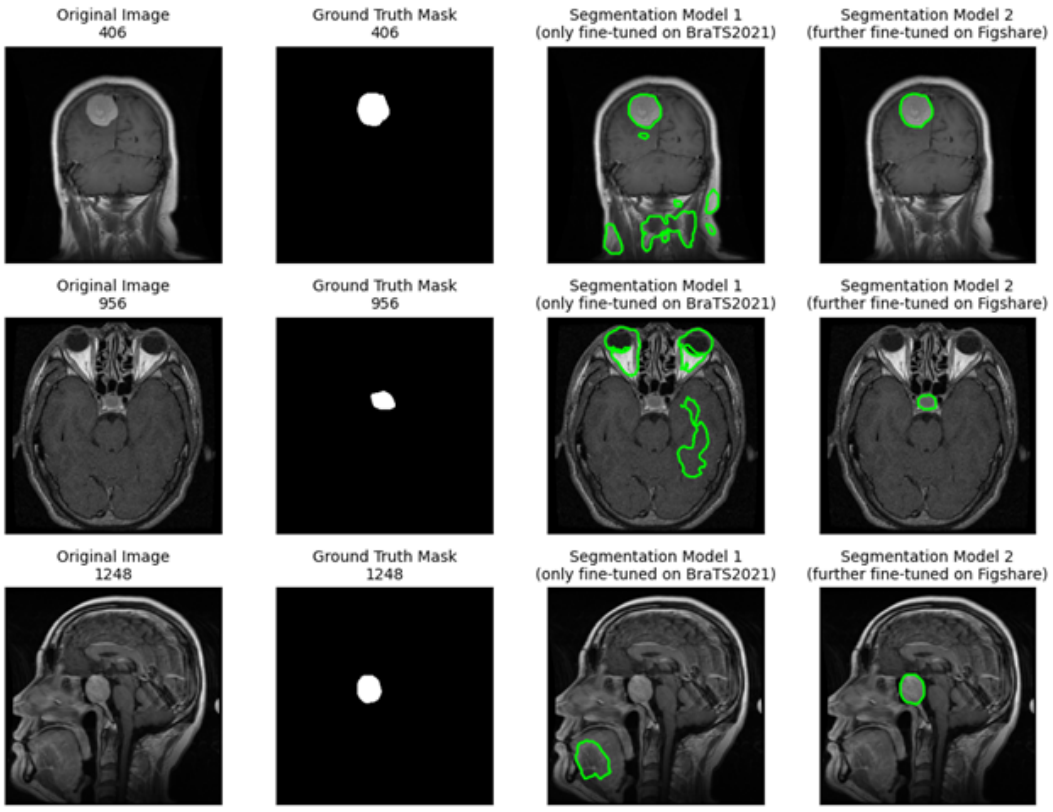
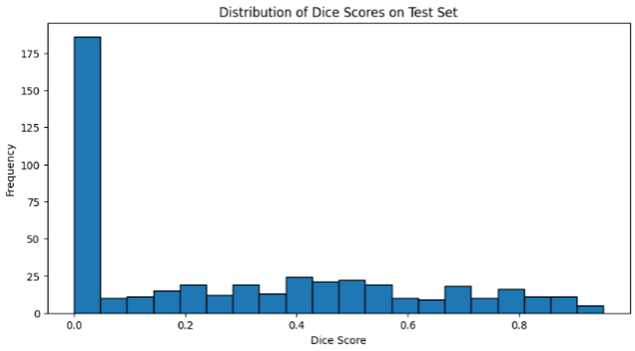
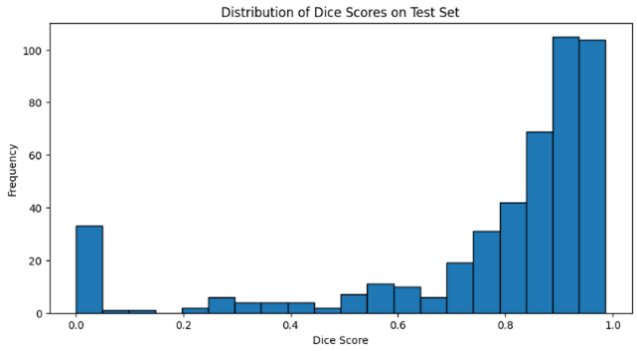


Figure 13: Comparison of brain tumour segmentation between predictions between BraTS2021-trained model, and Figshare fine-tuned model.

As shown in **Fig. 13**, further fine-tuning resulted in a drastic improvement in segmentation performance. This can be backed-up by assessing the Dice scores on the dataset, with a median of 0.2837 without further fine-tuning, and 0.7693 after additional fine-tuning, a stark difference. Furthermore, the distribution of Dice scores in **Fig. 14** shows a strongly skewed distribution for the former model, with an extremely high proportion of Dice scores around zero.



(a) Dice Distribution Histogram on Figshare Testing Dataset, with model not fine-tuned on figshare



(b) Dice Distribution Histogram on Figshare Testing Dataset, with model fine-tuned on figshare

Figure 14: Dice Distributions on Figshare Testing Dataset, (a) without further fine-tuning on Figshare, (b) with further fine-tuning.

The final output of the automated pipeline is presented in **Fig. 15**, where the tumour region is clearly marked with a green outline, and the predicted class, as well as its confidence score, is displayed. In cases where the no-tumour class is detected, no tumour outline will be displayed, and instead the system will just show the confidence score in the absence of a tumour. For further clarity, the confidence scores of all classes are revealed above in descending order, to allow clinicians to assess the level of certainty associated with each prediction, ensuring more informed decision-making.

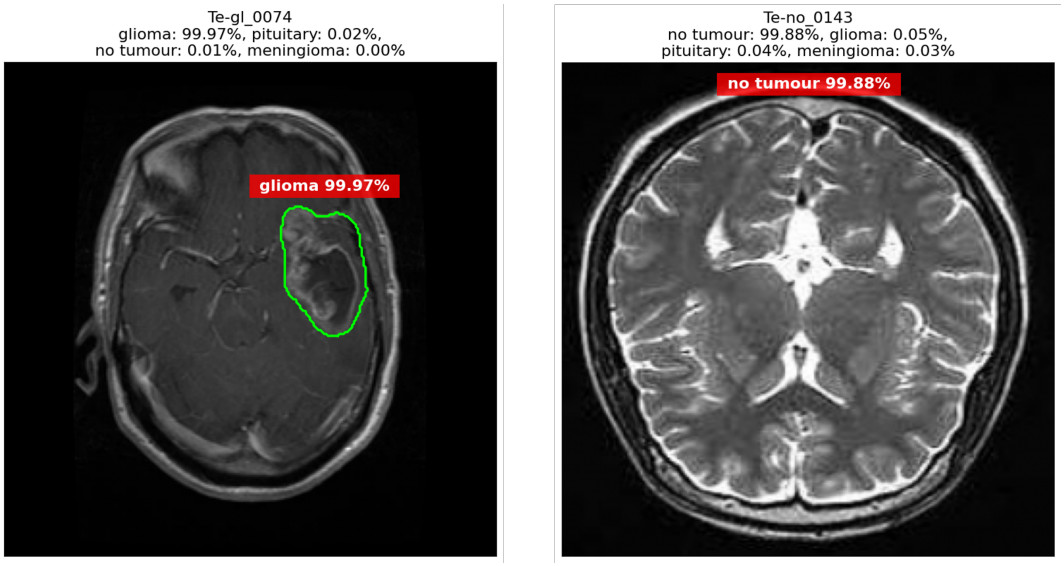


Figure 15: Outputs of pipeline, with Glioma image on the left and No-tumour on the right.

4 Project discussion, reflection and conclusion

4.1 Project Discussion

The goal of this project was to create an automated pipeline for brain tumour diagnosis on MRI scans, utilising both segmentation and classification tasks. The findings for both tasks demonstrate the strength and practicality of current deep-learning methods in enhancing the accuracy and efficiency of medical diagnoses. Through the use of pre-trained models, this project demonstrated the power of transfer learning and hyperparameter optimisation in improving model performance.

The classification model, based on the pre-trained GoogLeNet architecture, showed significant performance gains through transfer learning and hyperparameter optimisation. The base model showed poor performance, with an accuracy of just 32%. However, by leveraging transfer learning, the model's performance improved drastically to an accuracy of 94%, with increased accuracy of 98% after training with HPO. The HPO model not only achieved more accurate predictions, but also exhibited extremely high confidence in its predictions.

In segmentation, the baseline U-Net struggled due to differences in dataset characteristics, achieving a median Dice score of only 0.0995. Fine-tuning the model on the BraTS2021 dataset signficantly improved performance, with Dice scores rising to 0.77. Through further refinement, including the use of advanced loss functions and data augmentations, the final segmentation model achieved Dice scores of 0.81, demonstrating the value of optimised training and data augmentation.

One critical observation in this project was the impact of segmentation on classification. While the segmentation model demonstrated strong tumour detection, its use in classification led to a reduction in overall classification performance. This insight led to a shift in the original pipeline design, though it was beneficial to identify this issue early on in the process.

4.2 Reflection and Further Research

Reflecting on the outcomes of this project revealed several strengths and weaknesses. One key strength was the ability to leverage existing architectures in GoogLeNet and U-Net to create a robust and effective pipeline that aggregates classification and segmentation. The use of transfer learning was especially valuable, as it allowed the models to take advantage of pre-trained weights while reducing training time and computational resources required. Techniques such as data augmentations and fine-tuning on more diverse data elevated the robustness of the model.

For classification, the model achieved an accuracy of 98.34%, slightly outperforming previous research (98.04% on GoogLeNet) without the use of data augmentations. It's also comparable to VGG16's 98.69% accuracy [13], even though VGG16 is far more computationally complex. While newer models like Inception V3 [35] could potentially build on these results, the current performance is exceptional for this task.

Several future directions can be explored to further develop this project. Despite improvements in segmentation, the project still faced some challenges with small and irregular tumour shapes, with some dice scores close to zero. A median dice score of 0.81 in this work, though strong, shows that the model still has room for improvement, particularly when it comes to generalising to other datasets. When comparing these results to other studies, some report segmentation models achieving Dice scores over 0.90 with more advanced architecture, such as implementing a diffusion model to address systematic and random errors in segmentation masks [19]. U-Net models have proven to be highly effective at this task, but alternatively, the exploration of more advanced U-Net models, such as attention U-Net or 3D U-Net, could perform better. 3D U-Net [36] models would work seamlessly with the BraTS2021 dataset and

will enhance spatial information, but will be highly computationally demanding due to the added dimensionality of data. Additionally, while this project focused on T1-CE scans, some studies have employed multi-modal imaging to provide richer feature representations [15], but this would ultimately come down to the specific type of problem at hand. Another potential improvement to enhance clinical relevance would be to train segmentation on tumour sub-regions, such as the necrotic core or peritumoral edema, which can better quantify tumour structure and activity [18].

Another potential improvement is addressing the issue of overfitting. While slice-based variations, as well as training techniques like early stoppers were deployed, further steps could be taken to mitigate overfitting, such as employing data augmentation techniques such as biased crop, Gaussian noise and rotation. Moreover, although steps were taken to improve the clarity of predictions through confidence scores, additional steps can be taken to benefit the interpretability of results, particularly by clinicians whom may not have the deepest knowledge in ML and DL techniques. A growing area of research that is starting to make significant progress in the field of medical diagnostics is explainable AI (XAI). XAI, like Grad-CAM [37] could provide a more detailed analysis into which regions of the image influence predictions, and can build further trust and understanding among clinicians.

Finally, for a more efficient and fluid pipeline, YOLO-based joint architecture (such as YOLOv8 [38]) could be deployed for this task, as they are capable of both segmentation and classification simultaneously.

4.3 Conclusion

In conclusion, this project presents a robust and clinically relevant tool for the automated diagnosis of brain tumours using MRI scans. Through the use of gold-standard datasets, rigorous fine-tuning of deep-learning architecture, and thorough evaluation, the system demonstrates great performance in classifying and segmenting brain tumours, with 98.34% accuracy in classification and a Dice score of 0.81 for segmentation.

This project highlights the importance of data quality, model transparency, and contextual information to ensure the tool is not only accurate but interpretable and clinically viable. The project successfully achieved its initial objectives, and with access to additional resources and further research, the pipeline remains adaptable and ready for future enhancements.

Ultimately, the project contributes toward AI-assisted diagnostics, supporting more accurate, faster, and transparent decision-making in the treatment of brain tumours. With further research and deployment, it has strong potential for real-world application in healthcare settings.

References

[1] Jacques Ferlay, Isabelle Soerjomataram, Rajesh Dikshit, Sultan Eser, Colin Mathers, Marise Rebelo, Donald Maxwell Parkin, David Forman, and Freddie Bray. Cancer incidence and mortality worldwide: sources, methods and major patterns in globocan 2012. *International journal of cancer*, 136(5):E359–E386, 2015.

[2] The Brain Tumour Charity. Statistics about brain tumours. `https://www.thebraintumourcharity.org/get-involved/donate/why-choose-us/the-statistics-about-brain-tumours/`.

[3] NHS. Brain tumours. `https://www.nhs.uk/conditions/brain-tumours`, 2023.

[4] Wikipedia contributors. Nervous system tumor. `https://en.wikipedia.org/wiki/Nervous_system_tumor`, 2025.

[5] Aaron Cohen-Gadol. Brain tumor statistics, 2024. URL `https://www.aaroncohen-gadol.com/en/patients/brain-tumor/types/statistics`.

[6] The London Clinic. Benign brain tumours. `https://www.thelondonclinic.co.uk/services/conditions/benign-brain-tumours`.

[7] Wu Deng, Qinke Shi, Miye Wang, Bing Zheng, and Ning Ning. Deep learning-based hcnn and crf-rrnn model for brain tumor segmentation. *iEEE Access*, 8:26665–26675, 2020.

[8] Milica M Badža and Marko Č Barjaktarović. Classification of brain tumors from mri images using a convolutional neural network. *Applied Sciences*, 10(6):1999, 2020.

[9] Mahmoud Khaled Abd-Ellah, Ali Ismail Awad, Ashraf AM Khalaf, and Hesham FA Hamed. A review on brain tumor diagnosis from mri images: Practical implications, key achievements, and lessons learned. *Magnetic resonance imaging*, 61:300–318, 2019.

[10] Javier E Villanueva-Meyer, Marc C Mabray, and Soonmee Cha. Current clinical brain tumor imaging. *Neurosurgery*, 81(3):397–415, 2017.

[11] Himani Bhavsar and Mahesh H Panchal. A review on support vector machine for data classification. *International Journal of Advanced Research in Computer Engineering & Technology (IJARCET)*, 1 (10):185–189, 2012.

[12] Vijay Wasule and Poonam Sonar. Classification of brain mri using svm and knn classifier. In *2017 third international conference on sensing, signal processing and security (ICSSS)*, pages 218–223. IEEE, 2017.

[13] Arshia Rehman, Saeeda Naz, Muhammad Imran Razzak, Faiza Akram, and Muhammad Imran. A deep learning-based framework for automatic brain tumors classification using transfer learning. *Circuits, Systems, and Signal Processing*, 39(2):757–775, 2020.

[14] Jun Cheng. Figshare brain tumor dataset. `https://figshare.com/articles/dataset/brain_tumor_dataset/1512427`, 2024.

[15] Gopal S Tandel, Ashish Tiwari, and Omprakash G Kakde. Performance optimisation of deep learning models using majority voting algorithm for brain tumour classification. *Computers in Biology and Medicine*, 135:104564, 2021.

[16] Erena Siyoum Biratu, Friedhelm Schwenker, Yehualashet Megersa Ayano, and Taye Girma Debelee. A survey of brain tumor segmentation and classification algorithms. *Journal of Imaging*, 7(9):179, 2021.

[17] Spyridon Bakas. Rsna-asnr-miccai brain tumor segmentation (brats) challenge 2021. http://braintumorsegmentation.org/, 2021.

[18] Abdulkerim Duman, Oktay Karakuş, Xianfang Sun, Solly Thomas, James Powell, and Emiliano Spezi. Rfs+: A clinically adaptable and computationally efficient strategy for enhanced brain tumor segmentation. *Cancers*, 15(23):5620, 2023.

[19] Wenqing Li, Wenhui Huang, and Yuanjie Zheng. Corrdiff: corrective diffusion model for accurate mri brain tumor segmentation. *IEEE Journal of Biomedical and Health Informatics*, 28(3):1587–1598, 2024.

[20] Nahian Siddique, Paheding Sidike, Colin Elkin, and Vijay Devabhaktuni. U-net and its variants for medical image segmentation: theory and applications. *arXiv preprint arXiv:2011.01118*, 2020.

[21] Khiet Dang, Toi Vo, Lua Ngo, and Huong Ha. A deep learning framework integrating mri image preprocessing methods for brain tumor segmentation and classification. *IBRO neuroscience reports*, 13:523–532, 2022.

[22] Jana Fehr, Brian Citro, Rohit Malpani, Christoph Lippert, and Vince I Madai. A trustworthy ai reality-check: the lack of transparency of artificial intelligence products in healthcare. *Frontiers in Digital Health*, 6:1267290, 2024.

[23] Masoud Nickparvar. Brain tumor mri dataset. https://www.kaggle.com/datasets/masoudnickparvar/brain-tumor-mri-dataset/data, 2021.

[24] Nikhil Tomar. Brain tumor segmentation. https://www.kaggle.com/datasets/nikhilroxtomar/brain-tumor-segmentation, 2022.

[25] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9, 2015.

[26] mrgrhn. Googlenet (inceptionv1) with tensorflow. https://ai.plainenglish.io/googlenet-inceptionv1-with-tensorflow-9e7f3a161e87, 2021.

[27] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009. doi: 10.1109/CVPR.2009.5206848.

[28] Hugging Face. Googlenet - community computer vision course. https://huggingface.co/learn/computer-vision-course/unit2/cnns/googlenet.

[29] Jeremy Jordan. Common architectures in convolutional neural networks, 2018. URL https://www.jeremyjordan.me/convnet-architectures/.

[30] PyTorch Contributors. Crossentropyloss. https://pytorch.org/docs/stable/generated/torch.nn.CrossEntropyLoss.html.

[31] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[32] Adithya Prasad Pandelu. Day 48: Training neural networks — hyperparameters, batch size, epochs. https://medium.com/@bhatadithya54764118/day-48-training-neural-networks-hyperparameters-batch-size-epochs-712c57d9e30c, 2024.

[33] Neri Van Otten. Weight decay in machine learning and deep learning explained & how to tutorial. https://spotintelligence.com/2024/05/02/weight-decay/, 2024.

[34] Mateusz Buda, Ashirbani Saha, and Maciej A Mazurowski. Association of genomic subtypes of lower-grade gliomas with shape features automatically extracted by a deep learning algorithm. *Computers in Biology and Medicine*, 109, 2019. doi: 10.1016/j.compbiomed.2019.05.002.

[35] Ramazan İncir and Ferhat Bozkurt. Improving brain tumor classification with combined convolutional neural networks and transfer learning. *Knowledge-Based Systems*, 299:111981, 2024.

[36] Raghav Mehta and Tal Arbel. 3d u-net for brain tumour segmentation. In *International MICCAI Brainlesion Workshop*, pages 254–266. Springer, 2018.

[37] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017.

[38] Sumit Pandey, Kuan-Fu Chen, and Erik B Dam. Comprehensive multimodal segmentation in medical imaging: Combining yolov8 with sam and hq-sam models. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2592–2598, 2023.