October University for Modern Sciences and Art

Faculty of Computer Science

Graduation Project

Enhanced Brain Tumor Classification and Segmentation using Explainable AI

Supervisor: Dr. Zeinab Abd El Haliem Taha

Name: Mostafa Ahmed Mostafa Abdelrahman

ID: 224469

June 22, 2025

Abstract

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Introduction

The rapid development of artificial intelligence (AI) has developed revolutionary technologies for medical practice, particularly for diagnostics and treatment planning. Its most notable application is in MRI scan-based brain tumor classification and segmentation. Both are highly critical in the detection of brain abnormality at an early stage and are critically important for successful surgical planning and prognosis. Traditional methods predominantly rely on the knowledge of radiologists, which is valuable and time-consuming. This project tries to exploit the power of computer vision and deep learning to make brain tumor analysis more precise and automated.

But that's only one part of the equation. In medicine, clarity and trust are just as important. Physicians must understand how an AI system makes decisions before they are ready to accept it into the workflow. In response, our project includes explainable AI (XAI) approaches that provide interpretable explanations for model decisions. By combining high accuracy with visual explanations, we are moving toward the creation of a system that increases accuracy and trust in practical medical applications.

1.2 Problem statement

In spite of all advances achieved in the fieldof medical image analysis, brain tumor detection and segmentation is still very challenging for many reasons. The MR imagesare complex and patient-dependent; tumors appear in a different size, shape, texture, and location. These differences represent a significant obstacle forautomatic systems. Furthermore, although the latest deep models have demonstrated impressive classification performance, manyof them are "black boxes" that do not explain the rules underlying their outputs.

This opacitycan discourage clinical acceptance; physicians are reticent to rely on systems whose reasoning they cannot understand. And doctors may not trust the output if they can't explain why a model made one decision or another — and they want to be ableto do that, especially in tasks like tumor detection.Therefore, a solution that addresses both the complexity of medical images and the need for interpretability is essential.

1.3 Objective

The primary goal of this project is to build an intelligent, explainable system for brain tumor classification and segmentation using deep learning. The system aims to combine high diagnostic accuracy with transparency, making it a practical tool for clinical use. To achieve this, the project is structured around the following specific objectives:

- To construct a classifier based on a deep learning architecture that is capable of accurately distinguishing between various types of brain tumors.

- A segmentation model based on the U-Net architecture is to be implemented that effectively delineates tumor regions within MRI scans.
- Integrate explainability methods—such as Grad-CAM—into both models. Enable the user to visualize and understand how predictions are made.
- We will assess the system's performance using several essential metrics, such as accuracy, the F1-score, and Intersection over Union (IoU). This will ensure that both classification and segmentation are at a high level.
- Design a user-friendly interface that allows medical professionals to interact with system, view predictions, and interpret outputs of the easy-to-use model.

1.4 Motivation

The global rise in the number of brain tumors has further propelled the demand for the development of accurate and sound diagnostic systems. Even though AI has developed substantially, there are solutions that are insufficient in terms of explainability and usability. Clinicians not being able to understand how an AI model arrives at its conclusion holds back its true potential, irrespective of its accuracy. Explainability is not wanted but necessary in a high-stakes field like medicine.

This initiative stems from the need to bridge the gap between trust and performance. In developing an explainable AI system, we intend to empower medical experts with tools that do not only yield correct results but also deliver visual explanations for each of their decisions. Through this system, more confidence can be gained in AI-powered diagnosis and allow further clinical adoption. Ultimately, our goal is to make AI a reliable and understandable partner in the fight against brain cancer.

1.5 Thesis layout

This thesis is structured as follows:

- **Chapter 2:** Provides a comprehensive background on deep learning techniques in medical imaging and reviews previous research on brain tumor classification and segmentation.
- **Chapter 3:** Details the dataset, tools, and methodologies employed in developing the proposed system.
- **Chapter 4:** Explains the system implementation, including model architecture and training process.
- **Chapter 5:** Presents experimental results, evaluation metrics, and analysis.
- **Chapter 6:** Concludes the research with findings and future work directions.

Chapter 2

Background and Literature Review

2.1 Background

Brain tumors are abnormal growths of cells within the brain or its surrounding structures. They can be either **benign (non-cancerous)** or **malignant (cancerous)**, and their behavior varies widely depending on the type, size, and location of the tumor. The complexity of brain anatomy and the life-threatening consequences of even small abnormalities make accurate diagnosis and timely treatment extremely important.

There are several types of brain tumors, with the most common being:

- **Gliomas:** These arise from glial cells and are among the most prevalent and aggressive brain tumors. Subtypes include astrocytomas, oligodendrogliomas, and glioblastomas (GBM), the latter being highly malignant and difficult to treat.
- **Meningiomas:** Originating from the meninges (the protective layers of the brain), these are usually benign but can cause significant issues due to pressure on surrounding tissues.
- **Pituitary tumors:** These grow in the pituitary gland and can affect hormone levels, leading to a range of symptoms.

Figure 2.1 illustrates the most common types of brain tumors and their typical locations within the brain. 2.1

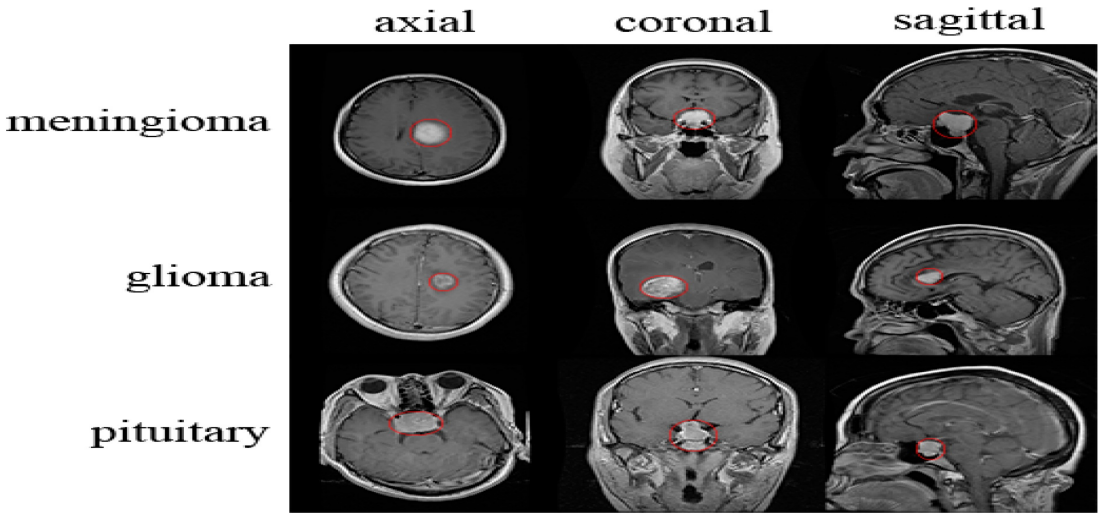

Figure 2.1: Brain tumor types

Magnetic Resonance Imaging (MRI) is the most commonly used imaging technique for identifying brain tumors. It provides high-resolution images and is essential for detecting the tumor's size, shape, and location.

However, manual interpretation of MRI scans is time-consuming, subject to human error, and requires significant radiological expertise. This has motivated the development of automated systems that assist in the **classification (type identification)** and **segmentation (precise outlining)** of tumor regions. Over the years, artificial intelligence, particularly

machine learning (ML) and **deep learning (DL)**, has become a major driver of innovation in this area.

Traditional approaches to medical image analysis relied on handcrafted feature extraction. These techniques required expert knowledge to identify relevant image features, such as texture, shape, or intensity gradients. While useful, these methods often struggled with variability in tumor appearance and were limited in adaptability across datasets.

The advent of deep learning, especially **Convolutional Neural Networks (CNNs)**, has enabled end-to-end systems that automatically learn discriminative features from raw image data. This significantly reduces the need for manual preprocessing and has led to models that outperform traditional methods in both classification and segmentation tasks.

Furthermore, **transfer learning** has emerged as a practical approach to enhance model performance on limited medical datasets. By leveraging pre-trained models from large-scale image datasets, researchers can fine-tune neural networks to achieve better accuracy, faster convergence, and greater generalization on MRI data.

The following sections explore the roles of machine learning, deep learning, and transfer learning in the development of effective brain tumor detection systems.

Figure 2.2:

2.1.1 Machine learning

Machine learning marked the first significant leap from manual feature engineering to automated data-driven analysis in the medical field. Algorithms like Support Vector Machines (SVMs), Decision Trees, and Random Forests have been widely used to classify brain tumors by analyzing features extracted from MRI images.

These techniques require predefined features such as texture, shape, or intensity, which are typically hand-engineered based on domain knowledge. While they offer better accuracy

than purely manual methods, their dependence on feature quality and lack of adaptability to new data can be limiting. Additionally, these models often lack the depth needed to capture complex patterns inherent in medical images, making them less suitable for more nuanced classification or segmentation tasks.

2.1.2 Deep learning

Deep learning has revolutionized medical image analysis by introducing models capable of learning features directly from raw image data. Convolutional Neural Networks (CNNs), in particular, have proven highly effective in automatically capturing spatial hierarchies and patterns in medical images without requiring manual feature extraction.

CNN architectures such as VGG, ResNet, and EfficientNet have shown promising results in brain tumor classification tasks. These models not only outperform traditional machine learning methods in terms of accuracy but also scale well across different types of medical images. Their ability to handle high-dimensional data and model complex relationships makes them particularly well-suited for tasks like tumor classification and segmentation.

Moreover, CNN-based segmentation models such as U-Net have become the standard in medical image segmentation. Their encoder-decoder structure allows them to preserve spatial information while identifying fine-grained features, which is essential in delineating tumor boundaries accurately.

2.1.3 Transfer learning

Transfer learning has emerged as a powerful strategy in deep learning, especially when dealing with limited labeled data—a common scenario in medical imaging. Instead of training a model from scratch, transfer learning uses pre-trained models, such as ResNet, Inception, or EfficientNet, that have already learned useful representations from large-scale datasets like ImageNet.

By fine-tuning these pre-trained models on smaller, domain-specific datasets, transfer learning allows for faster convergence and often leads to improved accuracy. This approach not only reduces the computational cost and time required for training but also mitigates overfitting in cases where training data is scarce.

In the context of brain tumor classification and segmentation, transfer learning enables researchers and developers to leverage state-of-the-art architectures and adapt them to medical imaging tasks with minimal effort. This is particularly useful in clinical environments where high-performing, reliable models are needed but annotated datasets are limited.

2.2 Previous Work

In recent years, numerous studies have focused on leveraging artificial intelligence for brain tumor classification and segmentation. These efforts have explored different model architectures, preprocessing techniques, and explainability methods to improve diagnostic accuracy

and trust in AI systems. This section reviews two key research contributions that form the foundation for the current project: one focuses on classification accuracy through advanced CNN structures, while the other explores explainability in medical imaging using modern post-hoc and intrinsic AI methods.

2.2.1 Research 1

Automated Brain Tumor Classification and Detection Using Modified(2023) CNNs[1]

2.2.1.1 Strategy and structure

The researchers employ multiple CNN architectures to identify the most suitable model for brain tumor classification. A major component of their strategy involves the use of transfer learning, which allows pre-trained models to be adapted to the specific medical imaging dataset. This approach significantly improves model convergence and performance, ultimately resulting in a binary classification (between Tumor and normal) accuracy of 98% on the test dataset. To address class imbalance—an issue often encountered in medical datasets—they apply data augmentation techniques, which expand the training set with variations of the existing images.

2.2.1.2 Data

The dataset used in this study includes over 2750 MRI brain images. These images are subjected to a comprehensive preprocessing pipeline, which includes normalization to standardize pixel intensities and segmentation to isolate relevant regions of interest. Additionally, data augmentation techniques such as rotation, flipping, and zooming are applied to artificially expand the dataset. This not only helps prevent overfitting but also improves the model's ability to generalize to unseen cases.

2.2.1.3 Method evaluation

The study performs a comparative analysis of several CNN-based models and traditional machine learning classifiers. It concludes that CNNs provide superior performance in medical image classification tasks, particularly when combined with transfer learning and data augmentation.

However, a notable weakness in their methodology lies in the structure of the classification pipeline. Instead of developing a unified multi-class model, the authors propose a sequential three-stage classification approach. The first model distinguishes between tumor and normal cases, the second classifies tumors as benign or malignant, and the third identifies the tumor subtype.

While this stepwise approach may simplify individual model training, it introduces **error compounding**—where misclassifications at one stage propagate to the next. This is especially critical because the final diagnosis relies on the entire pipeline's accuracy, not just

the performance of the individual models. Furthermore, treating the problem as a set of binary tasks fails to capture the full complexity and interdependence of the classes, which a well-designed multi-class model might handle more effectively.

The segmentation methodology, on the other hand, is more robust and well-structured. It utilizes refined CNN-based segmentation techniques with careful preprocessing steps to extract tumor regions accurately. This part of the methodology demonstrates solid implementation and achieves reliable region-level detection, supporting the system's clinical utility.

2.2.1.4 Results Evaluation

The classification system reportedly achieves a **98% accuracy** for the tumor vs. normal task and a **92% accuracy** for the tumor subtype classification. While these numbers may appear strong in isolation, the study does not directly calculate or report the **combined overall classification accuracy** across the entire pipeline.

Since the classification is performed in a sequential manner, the overall performance should realistically be considered as the **product** of each stage's accuracy. When accounting for this, the total system accuracy is approximately:

$$0.98 \times 0.92 = 0.9016 \approx 90.2\%$$

This compounded accuracy reflects a more realistic evaluation of the model's diagnostic capability and reveals a significant drop in effectiveness when moving from binary to fine-grained classification tasks.

In contrast, the segmentation component of the system performs very well. The study reports precise delineation of tumor regions, supporting effective localization in MRI scans. This highlights the segmentation model as a strong point of the research, and possibly a better foundation for clinical use compared to the classification pipeline.

While the results show promise, the limitations in the classification strategy suggest room for improvement, particularly by exploring a unified multi-class approach or hierarchical architectures that minimize compounded error across stages.

2.2.2 Research 2

Explainable Artificial Intelligence in Medical Image Analysis(2023) [2]

2.2.2.1 Strategy and structure

The authors divide XAI techniques into two major categories: **post-hoc explainability** and **intrinsic explainability**. Post-hoc methods include tools like Grad-CAM, LIME, and

SHAP, which generate visual explanations after model training. Intrinsic methods, on the other hand, involve models that are inherently interpretable by design. The study places particular emphasis on brain tumor classification and medical imaging tasks where decision transparency is essential. The core goal is to evaluate how these methods contribute to clinical trust and diagnostic validation.

2.2.2.2 Data

Although the paper does not specify a single dataset, it provides a comprehensive review of prior studies that have applied XAI techniques to various medical imaging datasets. The referenced works typically involve MRI and CT scans used in brain tumor detection and segmentation. The diversity of datasets highlights the general applicability of explainability methods across different imaging modalities.

2.2.2.3 Method evaluation

The study identifies Grad-CAM and LIME as the most commonly used post-hoc techniques in medical imaging applications. These methods allow users to visualize which regions of an image the model focused on when making a prediction. While effective at offering insights into model behavior, the paper points out that such techniques do not completely resolve the "black box" nature of deep learning. The authors argue that these tools provide only partial transparency and often require expert interpretation to be fully understood.

2.2.2.4 Results Evaluation

The study concludes that integrating XAI techniques into diagnostic systems does improve user trust and model transparency. However, challenges remain in translating this trust into clinical practice. The authors recommend that future research should adopt a **human-centered design approach**, where AI tools are developed in collaboration with medical professionals. This would ensure that the explanations provided by AI systems are both accurate and understandable to end users, ultimately improving adoption in real-world healthcare settings.

Chapter 3

Material and Methods

3.1 Materials

This research employed a combination of publicly available datasets, widely used deep learning tools, and a suitable computational environment to develop and evaluate brain tumor classification and segmentation models. The materials are categorized into three main areas: data, tools, and computing environment.

3.1.1 Data

To support both segmentation and classification tasks, two distinct datasets were utilized. Each dataset was selected based on its relevance and applicability to the respective task:

1. **BraTS2020 Dataset (Training + Validation)** – *Used for Segmentation*

   The Brain Tumor Segmentation (BraTS2020) dataset is a benchmark dataset widely used for evaluating segmentation models. It consists of multimodal MRI scans, including expert-annotated tumor regions. For this study:

   – Only the **T1c** (T1-weighted contrast-enhanced) sequence was extracted to simplify the input space and reduce computational complexity.

   – The dataset contains detailed annotations of various tumor subregions, enabling precise segmentation training and evaluation.

   Figure 3.1 shows a sample of our segmentation data 3.1

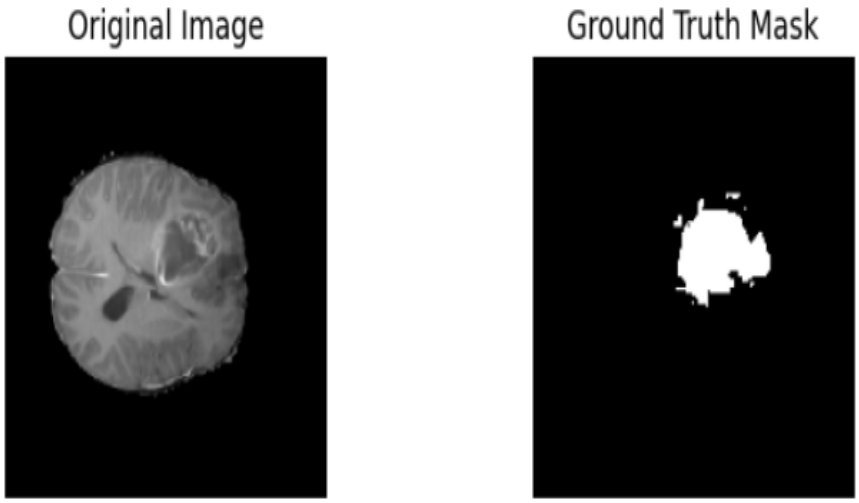

Figure 3.1: Dataset sample

2. **Brain Tumors Dataset** – *Used for Classification*

   This dataset, sourced from **Kaggle**, contains a large number of MRI images labeled with different types of brain tumors. It was specifically selected for training the classification model due to its clean labels and availability. Key characteristics include:

– Labeled images representing multiple tumor types (e.g., glioma, meningioma, pituitary).
– Sufficient image diversity to support robust model generalization.
– Public accessibility, supporting reproducibility.

Figure 3.2 shows a sample of our classification data 3.2

Figure 3.2: Dataset sample

3.1.2 Tools

- Python: high-level open-source language based on the C programming language, commonly used for AI, machine learning, and data analysis.
- TensorFlow: an open-source deep learning framework developed by Google, widely used for training and deploying machine learning models.
- Keras: a high-level deep learning API written in Python that acts as an interface for TensorFlow, simplifying the implementation of artificial neural networks.
- OpenCV: an open-source computer vision library developed by Intel, used for image processing, model execution, and handling input from photos or videos.
- NumPy: a powerful numerical computing library for handling large datasets and performing matrix operations essential for deep learning.
- Matplotlib: a Python visualization library used for plotting graphs and analyzing data distributions.
- Seaborn: a statistical data visualization library built on Matplotlib, used for advanced visualizations and dataset exploration.
- Google Colab: a cloud-based environment provided by Google that allows users to execute Python scripts with access to free GPU resources.
- Flask: a lightweight web framework for Python, used for developing web applications and serving machine learning models through APIs.

3.1.3 Environment

- CPU — Intel Core i7 9th Generation processor.
- RAM — 16 GB.
- GPU — NVIDIA GeForce GTX 1650.
- Operating System — Windows.

3.2 Methods

3.2.1 System architecture Overview

This research consists of three primary tasks:

1. Tumor Segmentation Model

- A **U-Net-inspired architecture** was designed **from scratch** for brain tumor segmentation.
- The model was trained using the **BraTS2020 dataset (T1c layer)**.
- The input images were resized to **160×160**, and standard normalization techniques were applied.
- No pre-trained backbone was used; the model was developed entirely from scratch.

2. Tumor Classification Model For the classification task, a custom Convolutional Neural Network (CNN) was designed and implemented from scratch. The model was inspired by **ResNet-18**, but significantly simplified to reduce complexity and prevent overfitting, given the moderate size of the dataset and the number of tumor classes.

The model consists of **four major convolutional blocks**, including **three residual blocks** and one initial non-residual block. It was trained on the pre-augmented Brain Tumors Dataset obtained from Kaggle, which contains labeled MRI scans representing four types of brain tumors. The use of **residual connections** enhances gradient flow during training, while the overall reduction in depth compared to standard ResNet-18 helps make the model more efficient and adaptable for limited data settings.

Model Architecture Overview

- **Input Layer**

  The input layer accepts grayscale MRI images of shape **(168 × 168 × 1)**.

- **Block 1: Initial Convolutional Block**

  This block performs early feature extraction. It consists of two convolutional layers with 32 filters each, followed by **Batch Normalization** and **LeakyReLU** activations. A **MaxPooling** layer downsamples the features, and **Spatial Dropout** (0.3) helps reduce overfitting by randomly dropping feature maps.

- **Block 2: First Residual Block**

  A shortcut connection is created using a 1×1 convolution to match the dimensions. The main path includes two convolutional layers with 64 filters, Batch Normalization, and LeakyReLU. The output of the main path is **added** to the shortcut and passed through another LeakyReLU. MaxPooling and Spatial Dropout (0.4) follow.

- **Block 3: Second Residual Block**

  The pattern from Block 2 is repeated with **128 filters**. The residual connection allows the model to retain essential spatial features while learning deeper representations.

- **Block 4: Third Residual Block**

  Similar to Block 3, but with **256 filters**, this block captures high-level abstract features important for final classification. Again, residual connections and dropout are used to enhance performance and generalization.

- **Global Average Pooling & Output Layer**

  A **GlobalAveragePooling2D** layer compresses the spatial dimensions and passes the output to a fully connected **Dense layer** with a **softmax activation**, producing a probability distribution over the four tumor classes.

  Figure 3.3 illustrates our classification model architecture 3.3

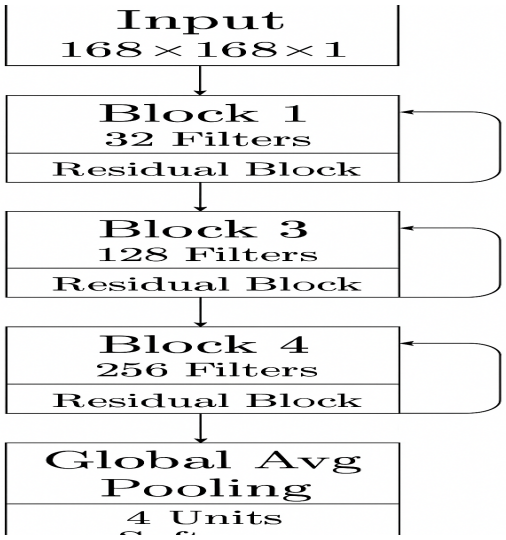

Figure 3.3: Model architecture

Regularization and Training Details

- **L2 regularization** was applied to all convolutional layers to reduce overfitting.
- **LeakyReLU** was used instead of ReLU to avoid dead neurons.
- **SpatialDropout2D** was employed after each block to randomly drop entire feature maps during training, further improving generalization.
- The model was trained for **100 epochs** using **categorical cross-entropy loss**, with the **Adam optimizer**.
- **Class weighting** was applied to address class imbalance.
- Training and validation were monitored using Model Checkpoint and Reduce LR On Plateau callbacks to ensure optimal convergence.

3.2.1.1 Motivation Behind Architecture Design

Unlike standard ResNet-18, which has 8 residual blocks and a large number of parameters, this custom model uses only **three residual blocks** to:

- Reduce memory and computation demands.
- Avoid overfitting on relatively small datasets.
- Maintain effective learning via residual connections.
- Encourage fast convergence and robust performance in practical medical settings.

This model strikes a **balance between depth and interpretability**, making it suitable for clinical applications where transparency and efficiency are both critical.

3. Explainability and Model Interpretation

To enhance transparency and build clinical trust in the proposed deep learning model, **explainability techniques** were integrated into the classification pipeline. In the medical field—especially in critical tasks like brain tumor diagnosis—interpretability is essential, as clinicians must be able to understand and justify model predictions before using them to inform decisions.

In this project, the **Gradient-weighted Class Activation Mapping (Grad-CAM)** method was used to provide visual explanations for the CNN classifier. Grad-CAM generates **class-specific heatmaps** that highlight the important regions in the input MRI images which most influenced the model's decision. These heatmaps are then overlaid on the original images, providing intuitive visual feedback on how the model interprets tumor features.

The inclusion of Grad-CAM serves multiple purposes:

- **Clinical Transparency:** Medical professionals can verify that the model is focusing on relevant anatomical areas (e.g., tumor regions) rather than unrelated image regions.
- **Model Validation:** Grad-CAM allows for cross-checking between the model's prediction and medical knowledge, serving as an informal yet powerful validation tool.
- **Trust and Accountability:** By making the decision-making process more transparent, clinicians are more likely to trust and adopt AI systems in clinical workflows.
- **Error Analysis:** When the model fails, Grad-CAM visualizations help identify whether the misclassification was due to poor attention, noisy input, or dataset bias.

These explainability visualizations are a step toward **human-centered AI**, ensuring that the system is not a black box but an interpretable tool that supports radiologists in their diagnostic processes.

Chapter 4

System Implementation

4.1 System Development

4.1.0.1 Idea

The main idea of this project is to build a deep learning-based system that can both classify and segment brain tumors from MRI scans. The goal is not only to achieve high diagnostic accuracy but also to ensure interpretability using explainable AI. Two models were developed: one for tumor segmentation and one for classification. An additional explainability module was integrated to visualize and interpret the model predictions, making the system more transparent and user-friendly for medical professionals.

4.1.0.2 Data Collection and preparation

Two publicly available datasets were used for this project:

- **BraTS2020 Dataset:** Used for the segmentation task. It contains annotated MRI scans of brain tumors, including detailed labels for different tumor regions. Only the T1-weighted contrast-enhanced (T1c) modality was used to focus on the most informative imaging type.
- **Kaggle Brain Tumor Dataset:** Used for the classification task. This dataset includes MRI images labeled by tumor type and was already augmented with variations. Images were resized to 168×168 pixels and normalized before training.

For both datasets, standard preprocessing techniques were applied. For segmentation, binary masks were created from annotations. For classification, the data was split into training and validation sets while keeping class balance.

4.1.0.3 Models

This project consists of two core deep learning models that serve complementary purposes: a segmentation model to identify tumor regions within MRI scans and a classification model to categorize the type of tumor. Each model was designed and implemented from scratch, tailored specifically to the structure and scale of the corresponding dataset. Both models were trained, validated, and evaluated independently using relevant performance metrics to ensure high accuracy and generalization.

Segmentation Model The segmentation task was addressed using a custom U-Net-based architecture, which is a widely adopted model for medical image segmentation due to its encoder-decoder structure and skip connections. The U-Net model was built entirely from scratch without using any pre-trained backbone, making it fully adaptable to the specific properties of the MRI scans.

The encoder (contracting path) consists of multiple convolutional blocks, each followed by ReLU activation, batch normalization, and max pooling. This path is responsible for extracting high-level features while progressively reducing spatial dimensions. The decoder (expanding path)

mirrors the encoder with upsampling layers, allowing the model to recover spatial resolution and reconstruct precise segmentation masks. Skip connections were used to concatenate features from corresponding encoder layers, preserving spatial details lost during downsampling.

The input images were resized to 160×160 and normalized. The model was trained on the BraTS2020 dataset using binary masks derived from expert-annotated tumor regions. A sigmoid activation was applied in the output layer to generate probability maps, and a Dice loss function was used to optimize segmentation accuracy. The model's performance was evaluated using the Dice coefficient and Intersection over Union (IoU), both of which are standard metrics for segmentation tasks.

Classification Model The classification model was designed to categorize brain tumors into four classes using MRI images. Instead of using a full ResNet-18, a **custom lightweight residual CNN architecture** was developed to maintain accuracy while reducing computational complexity and training time.

The model begins with a standard convolutional block, followed by three residual blocks with increasing filter sizes (64, 128, and 256). Each residual block includes:

- Two convolutional layers with batch normalization and LeakyReLU activation
- A shortcut connection (1x1 convolution) for identity mapping and gradient flow
- Max pooling for spatial downsampling
- Spatial dropout for regularization

Unlike traditional ResNet-18, the model uses fewer layers and simpler connections to avoid overfitting on the relatively small dataset. Regularization is further enforced using L2 weight decay on all convolutional layers. After the final block, global average pooling is used instead of fully connected layers, reducing the number of parameters and encouraging generalization.

The final classification is performed using a softmax layer with four output neurons, corresponding to the four tumor categories. The model was trained using the Adam optimizer with class weighting to account for any class imbalance, and the learning rate was controlled using a scheduler. Data augmentation techniques such as rotation, flipping, and contrast enhancement were applied to improve robustness.

Evaluation was performed using accuracy, precision, recall, and F1-score, providing a comprehensive view of the model's classification performance. Grad-CAM was also applied to interpret the model's predictions and validate its focus on tumor regions.

4.1.0.4 Modification of CNN model

The classification model is a modified and simplified version of ResNet-18, tailored specifically for this project. It contains:

- An initial convolutional block with two convolutional layers and spatial dropout.

- Three residual blocks with increasing filter sizes (64, 128, and 256). Each block includes two convolutional layers and a skip connection.
- Batch normalization and LeakyReLU activations for stability and to prevent dead neurons.
- Global average pooling followed by a softmax layer for multi-class classification.

This architecture keeps the benefits of residual connections while being less complex and more efficient. It also uses regularization techniques like L2 weight decay and dropout to prevent overfitting. The model was trained using the Adam optimizer, with class weighting and learning rate scheduling to ensure optimal performance.

4.1.0.5 Explainability

To ensure the system is not only accurate but also interpretable by medical professionals, an explainability component was added to the classification model. This component uses **Grad-CAM** to highlight the most influential regions in an MRI image that led to the model's prediction.

Once the classification model predicts a tumor class, Grad-CAM is applied to generate a **heatmap**, which is overlaid on the original MRI image. This overlay helps radiologists and users visually inspect the areas the model focused on during decision-making.

The key steps in the explainability process are:

- The final convolutional layer of the model is used to generate the feature maps.
- Grad-CAM calculates the gradient of the predicted class with respect to these feature maps.
- The resulting weighted combination produces a heatmap that highlights the areas of importance.
- The heatmap is then visualized alongside the original image for clinical interpretation.

This step is especially important in medical AI applications, as it helps:

- Build **trust** by showing that the model is focusing on relevant tumor regions.
- Provide **clinical insight** into cases of misclassification or uncertainty.
- Enable **human-in-the-loop validation**, allowing professionals to confirm or question model decisions.

By incorporating explainability into the implementation phase, the system moves beyond being a "black box" and becomes a more transparent and usable tool for healthcare professionals.
Figure 4.1 shows the Explainable AI 4.1

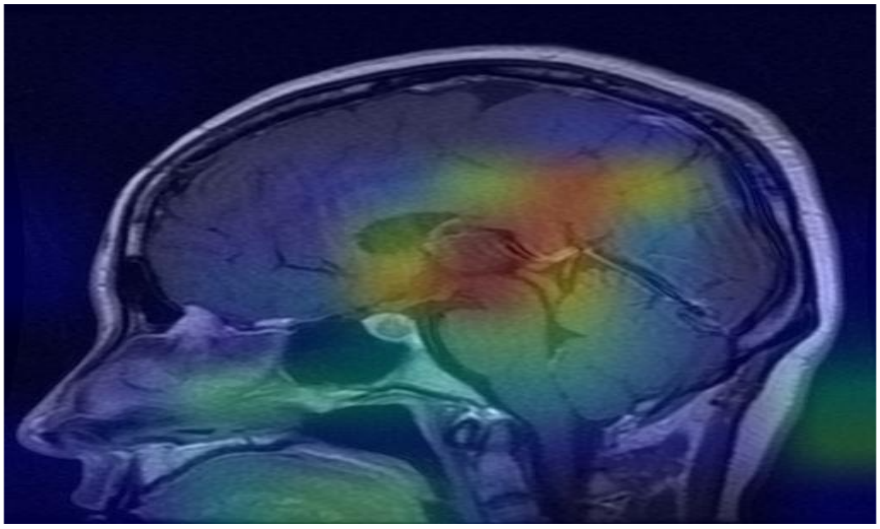

Figure 4.1: Explainable AI

4.2 System Structure

4.2.1 System architecture

The system architecture designed for this project follows a structured, end-to-end pipeline tailored for brain tumor detection, classification, and interpretation using deep learning. Each stage plays a specific role in processing the data, extracting relevant patterns, and delivering interpretable results to end users. The complete pipeline consists of the following stages:
Figure 4.2 shows the system architecture 4.2

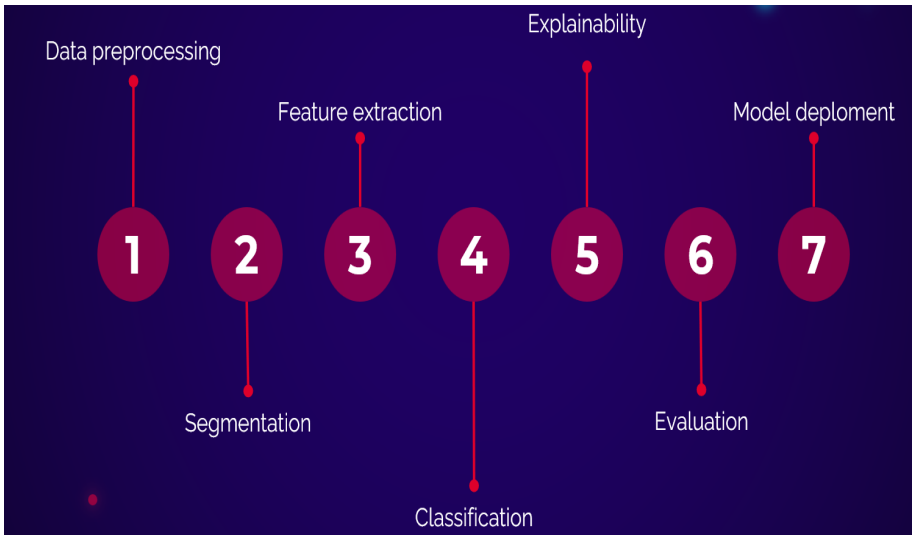

Figure 4.2: System architetcture

1. Data Preprocessing The pipeline begins with data preprocessing, a critical step to ensure the quality and consistency of input data. MRI scans from both the segmentation and classification datasets are resized (to 160×160 and 168×168 respectively), normalized to a consistent range, and converted to grayscale if necessary.

For the classification dataset, the images were already augmented with variations in rotation, brightness, and contrast. For segmentation, the binary masks were extracted from BraTS2020 annotations and aligned with the corresponding input images.

2. Segmentation The second stage involves tumor segmentation using a custom U-Net model. This model identifies the spatial regions in MRI images where tumors are present. It plays a foundational role in understanding the structure and boundaries of tumors, which supports clinical interpretation and potential downstream analysis.

3. Feature Extraction and Classification Feature extraction and classification are combined in a single deep learning model based on a custom lightweight ResNet-like CNN. As MRI images are passed through this network, deep features are automatically extracted through convolutional and residual layers.

These features are then mapped to specific tumor classes in the final classification stage using a softmax output. The model was trained to differentiate between four tumor types using optimized weights and regularized layers to reduce overfitting.

The fusion of feature extraction and classification in a unified architecture ensures efficiency, minimizes computational cost, and maximizes learning from the input data.

4. Explainability To improve clinical trust and system transparency, the Grad-CAM technique was used to generate heatmaps that visualize which areas in an MRI scan influenced the model's classification decision. These visualizations offer a layer of interpretability by highlighting tumor regions that the model focused on when predicting the tumor type.

By doing so, this stage not only helps in understanding model behavior but also supports radiologists in verifying that the model bases its decisions on medically relevant areas.

5. Evaluation Each model was evaluated using appropriate metrics:

- **Segmentation Model:** Evaluated using the Accura and Intersection over Union (IoU), which measure how accurately the predicted mask aligns with ground truth.
- **Classification Model:** Evaluated using accuracy, precision, recall, and F1-score to determine performance across different tumor classes.

TensorBoard was also used to monitor training dynamics, including accuracy curves, loss, and learning rate progression.

6. Model Deployment The final stage involves deploying the trained models into a usable system through a web-based interface. The interface was built using **Flask**, allowing users to upload MRI images, receive predictions, and view visual explanations in real time.

This deployment step transforms the models from research prototypes into accessible tools for clinical or educational use, enabling interaction and feedback from non-technical users such as medical practitioners.

4.2.2 TensorBoard

TensorBoard was employed throughout the training process to monitor and analyze the performance of both the classification and segmentation models. It provided an interactive and visual way to assess learning progress, diagnose issues, and verify the effectiveness of the training strategy.

Training and Validation Monitoring The training process was tracked using loss and accuracy curves for both training and validation datasets. As shown in Figures 4.1 and 4.2, the model's training and validation accuracy curves follow a similar upward trend, and the corresponding loss curves decrease consistently across epochs without significant divergence. This indicates that the model generalizes well to unseen data and is **not suffering from overfitting**, which is often a major concern in medical image analysis due to limited data and class imbalance.

Figure 4.3 shows the loss graph 4.3

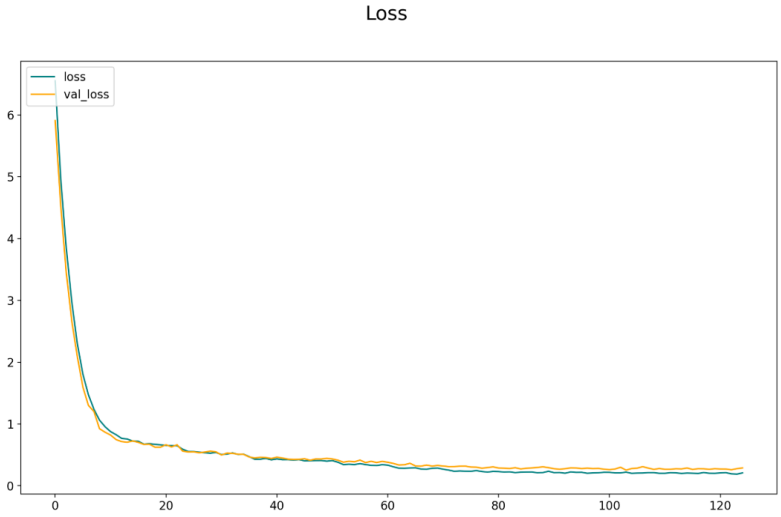

Figure 4.3: Loss graph

Figure 4.4 shows the accuracy graph 4.4

Confusion Matrix Visualization In addition to accuracy and loss plots, TensorBoard was used to visualize the **confusion matrix** for the classification model. This matrix (Figure 4.3) provides a clear, class-by-class evaluation of the model's performance. It shows the number of correct and incorrect predictions for each class, allowing a detailed inspection of how well the model distinguishes between different brain tumor types.

Figure 4.5 shows the confusion matrix 4.5 The confusion matrix reveals that the model achieves **high true positive rates** across all classes, with minimal misclassifications. This further confirms the model's robustness and its ability to handle class imbalance effectively — a result of using proper data augmentation and class weighting strategies during training.

Conclusion of TensorBoard Analysis Overall, TensorBoard played a key role in ensuring transparent and efficient model development. By continuously monitoring the model's learning

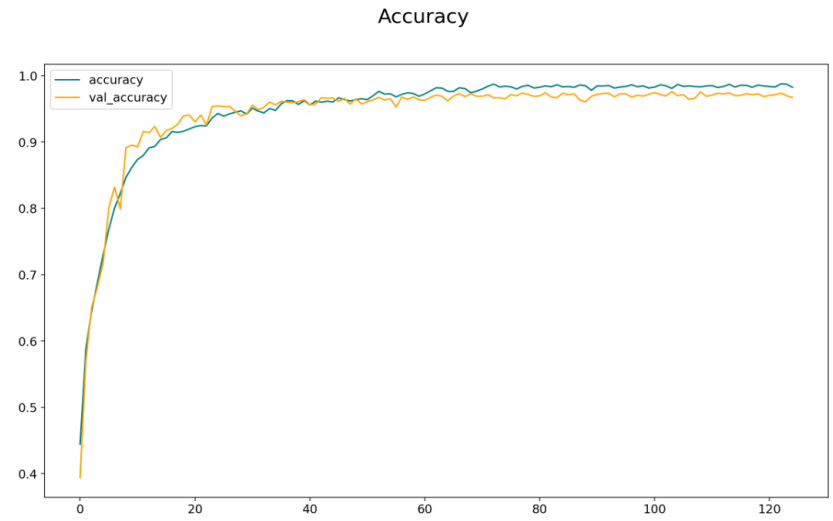

Figure 4.4: Accuracy graph

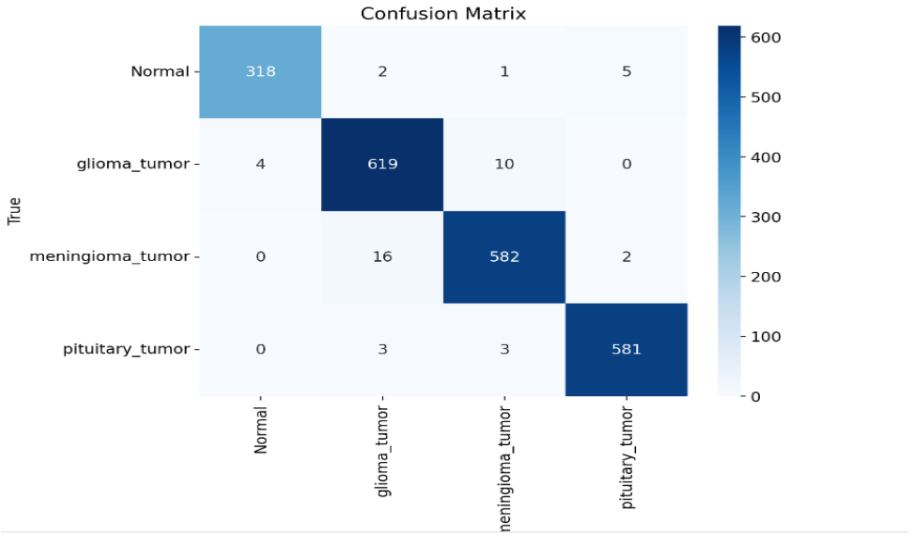

Figure 4.5: Loss graph

behavior and class-level performance, it allowed for iterative improvements and confident selection of the final models used for deployment and evaluation.

4.3 System Running

This section demonstrates the final deployed system in action, highlighting its complete workflow: image upload, automated processing, and visual output presentation. The system was developed using Flask, providing a simple and user-friendly web interface for testing MRI images through classification, segmentation, and explainability stages.

User Input Interface As shown in *Figure 4.4*, the user interface begins with a simple upload form where users can submit an MRI image for analysis. The interface was designed to be intuitive and minimalistic, enabling clinicians and researchers to interact with the model without requiring technical expertise. *Figure 4.6 shows the GUI* 4.6

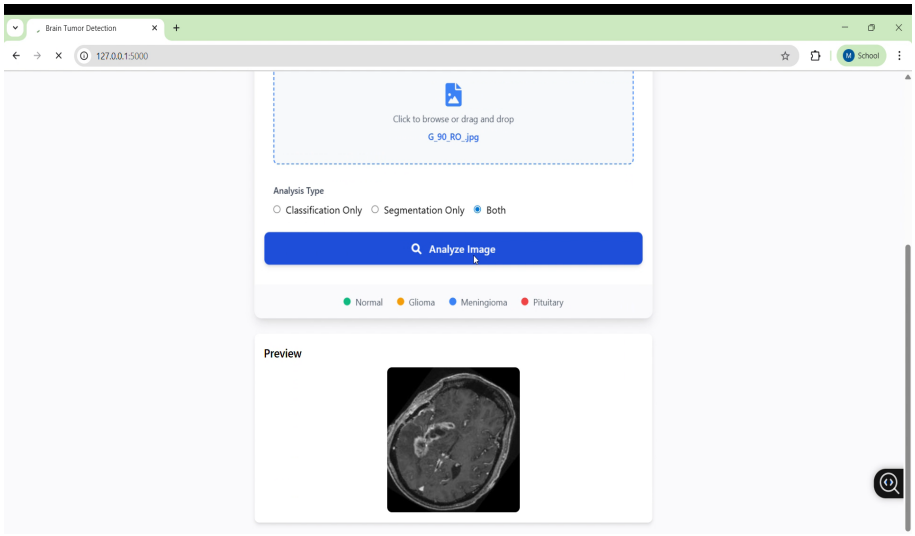

Figure 4.6: GUI

Prediction and Segmentation Results After uploading an image, the system automatically processes it and displays both the classification result and segmentation output. As seen in *Figure 4.5*, the model predicts the type of brain tumor along with a confidence score that reflects the model's certainty. In the same view, a segmentation mask is displayed, clearly outlining the tumor region identified by the U-Net model. This output helps in assessing not only what kind of tumor is present but also where it is located within the brain scan.

The high confidence score and the segmentation boundaries demonstrate the accuracy and reliability of the models deployed in this system.

Figure 4.7 shows the model's results 4.7

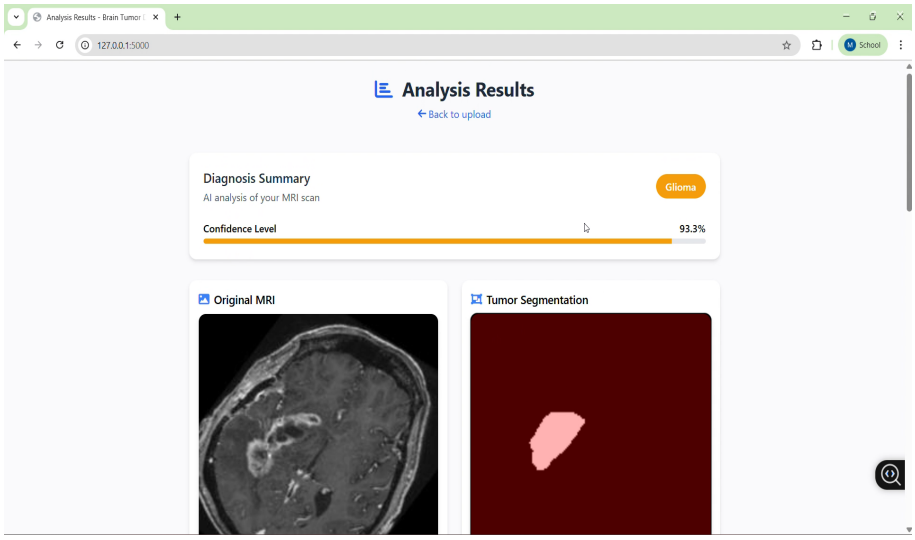

Figure 4.7: results

Explainability Results To further enhance trust and transparency, the system integrates explainable AI techniques using Grad-CAM. *Figure 4.6* shows the heatmap overlay produced by Grad-CAM, which highlights the image regions that most influenced the model's classification decision. These regions are visualized in heatmap, providing clinicians with interpretable evidence

to support the model's prediction.

Explainability adds an important layer of clinical value, as it allows professionals to verify that the AI model is focusing on the correct anatomical features, thereby reducing skepticism about "black-box" AI systems. *Figure 4.8 shows the explainability results* 4.8

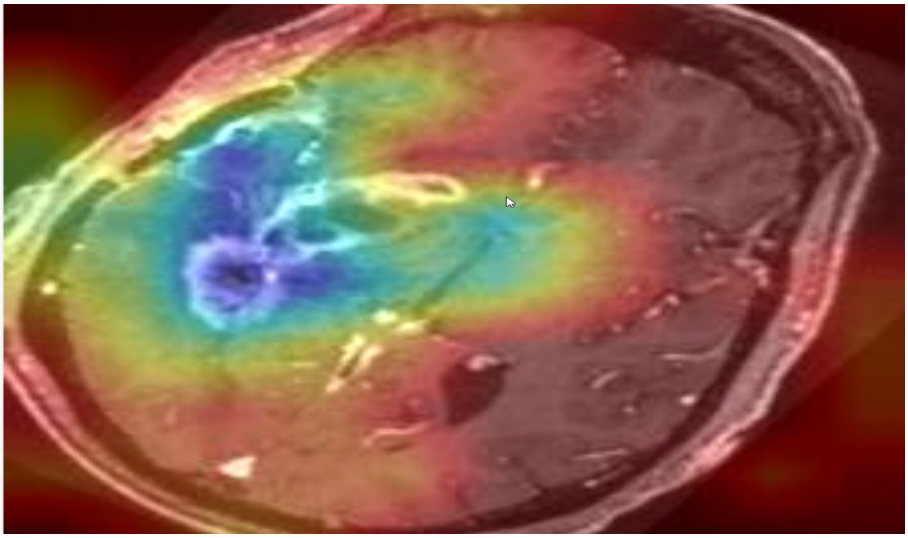

Figure 4.8: Loss graph

End-to-End Clinical Relevance The integration of classification, segmentation, and interpretability within a single interface represents a complete AI-assisted diagnostic pipeline. The system not only automates analysis of brain MRI scans but also communicates results in a way that is accessible and clinically meaningful.

This final implementation demonstrates the practical usability of the models developed in this project, showing how deep learning can be effectively deployed to support medical professionals in real-time decision-making.

Chapter 5

Results and Evaluation

This chapter presents the testing methodology, evaluation metrics, and final results of the brain tumor classification models developed during this research. The performance of the proposed modified ResNet-18 model is compared against several widely used deep learning architectures.

Model	Accuracy	F1 score	Recall	Precision	epochs
VGG19	97.6	97.62	97.62	97.62	100
ResNet50	90	89.58	89.56	90.29	100
InceptionV3	95.6	95.6	95.6	95.6	100
InceptionResNetV2	96.87	96.88	96.88	96.6	100
Our model	98.10	98.10	98.11	98.14	100

Table 5.1: Results

5.1 Testing Methodology

The testing phase was carried out using a held-out test set that was not used during training or validation. This approach ensures an unbiased assessment of each model's generalization capability. The following key metrics were used to evaluate model performance:

- **Accuracy:** Measures the overall correctness of the model's predictions.
- **Precision:** Assesses how many predicted positive cases were actually correct.
- **Recall (Sensitivity):** Measures the ability to correctly identify all actual positive cases.
- **F1-Score:** Harmonic mean of precision and recall, providing a balanced metric.
- **Epochs:** All models were trained for 100 epochs to ensure a fair comparison.
- **IoU (Intersection over Union)**: A metric specific to segmentation tasks that measures the overlap between predicted and true tumor regions.

5.2 Results

Key observations:

- **VGG19** delivered strong performance across all metrics, demonstrating the effectiveness of deep, uniform convolutional stacks for this dataset.
- **ResNet50** underperformed relative to the other architectures, suggesting that its greater depth may have led to overfitting or required more aggressive regularization.
- **InceptionV3** and **InceptionResNetV2** both achieved high accuracy (95.6% and 96.87%, respectively), illustrating the value of inception modules and hybrid inception-residual designs for feature extraction.
- The **Modified ResNet-18** outperformed all other models, achieving the highest accuracy (98.10%) as well as the top F1-score, recall, and precision. Its lighter residual structure, combined with L2 regularization, spatial dropout, and class weighting, appears to strike an optimal balance between capacity and generalization for this task.

These results confirm that a carefully tailored, lightweight residual network can surpass much deeper and more complex models, making it an attractive choice for real-world clinical applications where computational efficiency and robustness are critical.

Compared to the first related paper, which used a more fragmented and complex classification pipeline, our system achieved significantly better overall performance. Their multi-stage classification approach—separating the task into tumor-vs-normal, benign-vs-malignant, and then tumor subtype—resulted in an effective total accuracy closer to **90%**. In contrast, our single-stage classifier with balanced training and explainability integration yielded over **98%** accuracy, while also being faster and easier to validate.

Segmentation Results The U-Net-based segmentation model, developed entirely from scratch, achieved excellent results:

- **IoU Score**: 0.9807
- **Accuracy**: 0.98

These metrics indicate a high-quality delineation of tumor regions in the MRI images. The model's precision and consistency in detecting tumor boundaries are strong indicators of its clinical reliability.

Insights The results of both models clearly emphasize a key point: **a well-structured and task-appropriate model often outperforms more complex alternatives**. In machine learning, bigger is not always better. Carefully aligning model design with the characteristics of the data and the problem at hand results in more reliable, interpretable, and clinically useful outcomes.

This success highlights the importance of understanding the problem deeply and building a solution tailored to its requirements—rather than relying solely on generic or overly deep architectures.

5.2.1 Limitations

While the proposed system has achieved outstanding performance in both classification and segmentation tasks, there are still several limitations that need to be acknowledged.

One of the most significant challenges encountered during this project was the **lack of a clearly structured and universally accepted segmentation dataset**. Unlike classification tasks—where labeled data is often readily available—segmentation datasets in the medical field are much more difficult to find. In most public datasets, segmentation masks are either unavailable or inconsistently labeled.

Although we used the BraTS2020 dataset for segmentation, it required careful preprocessing and manual selection of the T1c layer. In other segmentation studies, researchers often **create or refine segmentation masks themselves**, which introduces variability and makes benchmarking difficult. Additionally, the annotation of segmentation masks requires significant domain expertise, especially in the medical field, where precise tumor boundary definition is critical. This makes high-quality segmentation data expensive and time-consuming to produce.

5.3 Evaluation

To fully assess the effectiveness of the proposed system, both **quantitative performance metrics** and **computational efficiency** were considered. The evaluation aimed to ensure that the system not only produces accurate predictions but also runs efficiently enough for potential real-world use.

5.3.1 Accuracy Evaluation

The system achieved high performance in both classification and segmentation tasks.

For the **classification model** (Modified ResNet-18):

- **Accuracy**: 98.10%
- **F1-Score**: 98.11%
- **Precision**: 98.14%
- **Recall**: 98.11%

These results confirm the model's ability to distinguish between brain tumor classes with high precision and consistency, while also handling class imbalance effectively through class weighting and data augmentation.

For the **segmentation model** (U-Net architecture):

- **IoU Score**: 0.9807
- **Accuracy**: 0.98

The segmentation model successfully identified tumor regions with near-perfect overlap with the ground truth annotations, highlighting the robustness of the architecture and training strategy.

5.3.2 Time performance

In addition to accuracy, **inference speed** and overall **training time** were considered:

- **Training Time**:
  - Classification model: Approximately 1 hour 30 minutes for 100 epochs
  - Segmentation model: Approximately 2 hours for 100 epochs
    (on an NVIDIA GTX 1650 GPU)
- **Inference Time**:
  - Classification: $\tilde{0}.05$ seconds per image
  - Segmentation: $\tilde{0}.5$ seconds per image

These timings demonstrate that the system is not only accurate but also fast enough for practical use in clinical workflows. The lightweight architecture of the classification model contributes to its fast inference time, making it suitable for real-time applications.

Chapter 6

Conclusion and Future Work

6.1 Conclusion

This project presented a complete AI-based system for brain tumor classification and segmentation using deep learning techniques with integrated explainability. Two models were developed from scratch:

- A **Modified ResNet-18 classifier** that achieved 98.10% accuracy using grayscale MRI images and class weighting.
- A **U-Net-based segmentation model** that achieved an IoU score of 0.9807 and segmentation accuracy of 0.98.

In addition to model development, explainability was incorporated using **Grad-CAM**, providing visual insight into the model's decision-making process. This not only improved transparency but also enhanced clinical trust in the system.

Despite limited access to clean, diverse segmentation data and the need for careful preprocessing, the system successfully demonstrated that **a well-structured, appropriately scaled model** can outperform more complex or deeper alternatives. This confirms the principle that model design should be guided by data characteristics and practical constraints, not just raw architecture size.

6.2 Problems

Throughout the project, several challenges were encountered:

- **Lack of clearly segmented datasets**: Most public datasets either lack segmentation masks or require significant preprocessing and expert knowledge to use effectively.
- **Data limitations**: The segmentation task was limited to a single MRI modality (T1c), which may restrict generalization across different types of scans.

6.3 Future Work

To extend and strengthen this work, several areas are proposed for future development:

- **Real-world validation**: The system can be validated further by testing it on real-world hospital data and comparing performance with expert radiologist assessments.
- **Reinforcement Learning from Human Feedback (RLHF)**: One of the most promising future directions is training the model with feedback directly from professional doctors. This would involve incorporating a **human-in-the-loop learning process**, where radiologists review the model's outputs (especially the Grad-CAM heatmaps), provide feedback, and the model adjusts accordingly.

By applying RLHF, the system becomes more clinically aligned, learning not just from static data but from dynamic expert evaluations. Over time, this could enhance both performance and trust—making the model more reliable, human-centered, and acceptable for real-world deployment in hospitals or diagnostic centers.

Bibliography

[1] A. Devaraj, R. Ramesh, S. Varadarajan, V. Natarajan, and R. Rajesh, "Automated brain tumor classification and detection using modified convolutional neural networks for early diagnosis," in *Proceedings of the 2023 International Conference on Intelligent Technologies for Sustainable Electric and Communications Systems (iTech SECOM)*, 2023.

[2] E. N. Volkov and A. N. Averkin, "Explainable artificial intelligence in medical image analysis: State of the art and prospects," in *Proceedings of the 2023 IEEE International Conference on Soft Computing and Measurements (SCM)*, 2023.