ID N. 16904

UNIVERSITÀ CAMPUS BIO-MEDICO DI ROMA

DEPARTMENT OF ENGINEERING

ISTITUTO NAZIONALE DI FISICA NUCLEARE

Italian National Ph.D. in Artificial Intelligence
Health and Life Sciences
XXXVII Cycle

An xAI Pipeline for MRI image segmentation for brain tumour identification: from model to GUI through ethics and European regulations

Supervisors

Cecilia Voena

Stefano Giagu

Candidate

Greta Grillo

May, 2025

To my family and friends.

Abstract

This doctoral thesis presents the design, implementation and validation of an engineering pipeline for explainable artificial intelligence (XAI) in the context of brain tumour segmentation from magnetic resonance imaging (MRI). The proposed system is explicitly developed in accordance with human-centered and regulatory principles—including user technological proficiency, accountability, transparency, completeness, and information format—aligned with current European AI guidelines.

The pipeline integrates state-of-the-art segmentation models with explainability tools, notably Grad-CAM, which is operationally embedded to provide visual saliency maps. In addition, the TracIn algorithm was studied and evaluated as an alternative attribution method and has been published in a separate work, although it was not integrated into the pipeline itself.

A modular graphical user interface (GUI) was developed to enable clinicians to interactively explore segmentation results, access algorithmic justifications, and interpret outputs using domain-specific visual tools.

A key innovation lies in the integration of GPT-4o, a multimodal large language model, to generate human-readable textual explanations directly from visual segmentation outputs. The system was evaluated through a qualitative study focused on zero-shot prompting, assessing GPT-4o's capacity to contextualize visual evidence without relying on traditional classification accuracy or quantitative interpretability metrics, and then tested with minimal modification of prompting to identify critical issues and improvements.

Experimental validation was performed using the Br35H and BraTS19 datasets. The findings demonstrate the potential of multimodal LLMs, while also reinforcing the importance of human oversight in medical decision-making. Finally, the thesis addresses legal and ethical considerations for deploying high-risk AI systems in healthcare, offering a replicable, regulation-aligned model for transparent and explainable AI in medical imaging.

Contents

List of Figures

List of Tables

List of Algorithms

- **Algorithm 11** – *ResUNet (Residual U-Net)*
  Implemented in Section 2.4.2 as an alternative segmentation architecture.

- **Algorithm 12** – *Attention U-Net*
  Also presented in Section 2.4.2, leveraging attention mechanisms for enhanced segmentation.

- **Algorithm 13** – *Softmax Normalization for pixel-wise classification*
  Employed in Section 2.4.2 during inference of segmentation masks.

- **Algorithm 14** – *Dice Loss function*
  Used in training (Section 2.4.2) to optimize tumour segmentation with imbalanced class distribution.

Chapter 1

Introduction

Artificial intelligence (AI) is profoundly changing the operational paradigm of contemporary medicine. In this context, Explainable Artificial Intelligence (XAI) emerges as a response to the growing need to understand, justify, and evaluate algorithmic decisions, especially when decisions impact patient health. The field of XAI has gained prominence in the medical field to mitigate the so-called "black-box effect" of deep learning models, which, while achieving high performance, do not offer explanations that are easily interpreted by clinical specialists. XAI , until about 5 years ago, was articulated and focused on two main methodological approaches: ante-hoc, in which interpretability is built into the structure of the model itself; and post-hoc, in which the explanation is derived a posteriori, once the model has been trained. Post-hoc techniques are the most widely used in clinical settings and include methods such as LIME, SHAP, CAM, LRP and Grad-CAM, which are capable of visualizing the areas of the image that most influence the model decision [1].
However, the widespread adoption of these techniques does not equate to their real understanding by clinicians, nor does it automatically guarantee trust and acceptance [2]. The actual concept of "explainability" is subject to varying interpretations: in the technical domain it refers to metrics of internal model consistency, while in the medical domain it translates into the ability to justify the decision in clinically relevant and useful terms. This diversification of meaning led to differing approaches to the problem, but also to heterogeneity of studies and thus to difficulty in pursuing a path to the goal. One of the main attempts to systematize the XAI landscape has been provided by Graziani et al. [3], who propose a comprehensive taxonomy of interpretable artificial intelligence. They distinguish between epistemic explicability (how the model works) and ethical explicability (who is responsible and how action can be taken). Specifically, epistemic explainability is understood as the system's ability to provide knowledge about its internal workings: it thus concerns the intelligibility of decision-making, the transparency of the model's structure, and the ability to

understand input-output relationships through formal or visual tools. Graziani et al. classify this dimension according to different levels of granularity, from the simple visualisation of activated features to the complete traceability of decision flows in complex models. Ethical explainability, on the other hand, is associated with the responsibility and social impact of the system: it does not merely clarify how the model produces an output, but implies the need to ensure that this output is justifiable, contextualised and reconstructible in the event of damage or error. In this sense, ethical explainability incorporates aspects of accountability and contestability, i.e. the possibility that a human user may question or correct an automated decision. Thus, in this context, a mapping that links technical methods to different types of human and social needs is proposed, emphasising that a unified approach to XAI must consider not only the technical functioning of the model, but also its inclusion in a regulated socio-technical ecosystem that is sensitive to fundamental human values.

1.1 Human-centered XAI and the clinical interface

According to these premises, in the healthcare context, explainability cannot be conceived as a purely technical attribute of the artificial intelligence model. Rather, it must be interpreted as a relational property, which depends on the interaction between the output of the system and the skills, expectations and needs of the physician. In other words, an explanation of a decision is only such if it is understandable and useful to the end user. This "user-centred" perspective is now central to the XAI debate and has led to the emergence of conceptual models capable of guiding the design of truly assistive systems.

One of the main limitations of current XAI systems is the lack of alignment between the content of the explanation and the mental model of the clinician[1]. The clinician, in fact, is not interested in knowing which neurons are active or which features are most influential in an abstract sense but rather wants to understand whether the algorithm has followed a coherent clinical reasoning and whether it has made recognisable or avoidable errors. Hence the need to develop evaluation metrics that are not limited to the fidelity or completeness of the output, but that take into account its usability in real-world contexts.

A second critical element concerns the level of technological literacy of the user. Wixom and Todd [4], integrating the technology acceptance model (TAM) with user satisfaction indicators, show how the perceived effectiveness of a system is mediated by the user's skill to interact with it, to understand its language and to integrate its results into their workflow. In the clinical setting, this translates into the need to design intuitive interfaces, which use a non-technical language and which allow the physician to validate or challenge the system's

decisions. Another critical issue is the method by which the explanation is presented. Visual explanations such as heatmaps can be useful, but are often ambiguous, while NLP-generated textual explanations[1] can introduce bias or over-interpretation.
Therefore, a multimodal approach should be adopted, in which different information formats (textual, visual, numerical) are integrated to answer different clinical questions. Finally, the issue of accountability is central: a system that provides explanations must also allow for traceability of decisions, so that in the event of an error, it is possible to trace it back to the source of the problem.
This is an essential requirement from both a medical and regulatory point of view, especially in high-risk contexts such as the diagnosis of brain tumours. Nevertheless, recent studies have highlighted that explainability, while necessary, is not always sufficient to ensure safe and effective clinical integration. The way in which the explanation is formulated and presented to the user plays a crucial role in its actual usability. Some authors propose structured evaluation frameworks for XAI systems that include criteria such as clarity, consistency, cognitive fit and presentation modality, underlining that explainability must be assessed not only for technical soundness but also for its interpretive effectiveness [5]. At the same time, the presence of an explanation module does not automatically imply that the physician will understand or use the information it provides. Experimental evidence shows that medical professionals often follow the AI's recommendation even when explanations are present, and even when those recommendations are flawed [6]. This behavioural tendency raises concerns about the true level of comprehension and interaction, pointing to the risk of over-reliance on systems perceived as authoritative.
Moreover, several critical voices in the literature have warned that current XAI approaches may offer only an illusion of transparency. Post hoc explanations, while seemingly plausible, may not reflect the actual internal reasoning of the model. In this sense, explainability risks becoming a superficial legitimisation layer, producing what some scholars call ersatz understanding—a simulated form of comprehension that does not improve accountability nor supports critical judgment [7].This is especially problematic in healthcare, where the stakes are high and blind trust in automated systems may lead to harmful consequences. Effective explainability in medicine cannot be left to the algorithm alone, but must emerge from a continuous interaction between machine and user, supported by human-centred designed interfaces validated in real clinical contexts.

[1]NLP : Natural Language Processing

1.2 The regulatory and ethical landscape

During the years of development of this thesis, the European Union has also made significant strides in regulating artificial intelligence, especially in high-risk contexts such as healthcare. The General Data Protection Regulation (GDPR), which came into force in 2018, already introduced the principle of the "right to explanation" for automated decision-making processes, establishing an initial regulatory reference to algorithmic transparency. However, it is with the approval of the Artificial Intelligence Act (AI Act), in its final version of 2024, that a systematic legal framework for high-impact AI systems is consolidated. The new regulation explicitly classifies AI systems used in the medical field as "high-risk systems", imposing stringent obligations in terms of safety, governance, human oversight, life-cycle management and above all explainability. Developers are required to document in detail the logic, inputs, intermediate and final decisions of the system, ensuring traceability and verifiability of algorithmic decisions [8].

At the same time, the ethical debate on the role of AI in society has grown stronger. In particular, the proposals developed by the AI4People group, led by Luciano Floridi, have contributed to structuring a shared value framework based on five principles: beneficence, non-maleficence, autonomy, justice and explicitness [9]. The latter principle, introduced as an original contribution of European thinking, emphasises the importance of providing answers that are not only technically correct, but also understandable, contextualised and accountable. The principle of explicability has two components: intelligibility (how the system works) and accountability (who is responsible for its functioning). According to Floridi and Cowls [10], this distinction is crucial in order to distinguish technical transparency from ethical and normative transparency. In healthcare, this implies that any decision proposed by an AI system must not only be understandable to the physician, but also justifiable in the event of a clinical-legal dispute. Another important initiative is the "Rome Call for AI Ethics" (2020), jointly promoted by the Pontifical Academy for Life, Microsoft, IBM and FAO. This document enshrines the commitment to an anthropocentric artificial intelligence that promotes human dignity, inclusion and sustainability. The document emphasises the need to empower developers and users of AI systems by promoting interdisciplinary dialogue between engineers, lawyers, physicians, patients and ethicists [11]. These guidelines, although geared towards development in the real environment, also influence the approach of the research and studies carried out by us researchers.

1.3 Thesis scope and application context

In this thesis, ethical and regulatory guidelines are not treated as external constraints but as foundational principles informing the system's design. The entire XAI pipeline was developed in accordance with the AI Act and the core values of explainable and responsible artificial intelligence. From dataset selection and data traceability to the design of the graphical interface and interactive language components, every aspect was conceived to promote transparency, human oversight, and respect for clinical autonomy. In this light, ethics is not merely a philosophical stance but a functional element embedded in both the technological and design process. Aligning technical performance with normative values is what ultimately distinguishes a reliable, adoptable, and socially legitimate system.

Building on this framework, the thesis addresses a specific and high-risk clinical application: the segmentation of brain tumours in Magnetic Resonance Imaging (MRI). MRI is a widely adopted, non-invasive imaging modality in neuroradiology, valued for its ability to produce high-resolution anatomical visualizations. Automated segmentation of these images can assist clinicians in detecting and delineating pathological areas, supporting diagnosis, treatment planning, and follow-up. However, the integration of deep learning models into such a critical domain poses substantial challenges in terms of interpretability.

This work presents the design, implementation, and experimental validation of an explainable AI pipeline for brain tumour segmentation in MRI. The system integrates a UNet-based segmentation framework with Grad-CAM for visual attribution and includes a clinician-oriented GUI for interactive exploration and validation of results.

While TracIn—a method for evaluating the influence of training data on predictions—was studied and resulted in a separate publication, it was neither integrated into the main pipeline nor used in the experimental validation involving large language models.

A central contribution of this thesis is the integration of GPT-4o as a multimodal explanation generator. Beyond qualitative assessments—which remain essential for responsible use—this thesis also employed quantitative evaluation metrics, including accuracy, precision, recall, and confusion matrices. These were used to assess the consistency and correctness of GPT-4o's outputs across various zero-shot prompting strategies. Experiments were conducted on the BR35H and BraTS19 datasets, aiming to evaluate the model's ability to detect, describe, and localise tumour features from visual input alone, and to investigate how prompt design influences its interpretive accuracy.

Overall, this thesis contributes a practical and modular framework that integrates ethical, legal, and technical dimensions, providing a replicable model for the transparent and trustworthy adoption of XAI in clinical workflows.

1.4 Comparison of Segmentation Architectures and Motivation for nnU-Net

Automatic segmentation of multimodal brain tumors in this thesis was addressed by implementing three state-of-the-art variants derived from the classical U-Net: **DeepSeg**, **ResUNet**, and **Attention U-Net**. These architectures share the encoder–decoder paradigm typical of U-Net, leveraging skip connections to preserve spatial information and combine local and global features during reconstruction [12].

DeepSeg, winner of the BraTS 2022 Challenge, stands out for its optimized pipeline combining batch normalization, targeted data augmentation, and efficient handling of both local and global features [13]. The refined skip connections reduce detail loss at tumor boundaries, improving the segmentation of structures such as peritumoral edema with fuzzy margins. In the validation set, DeepSeg achieved a Dice score of **0.87** for the Enhancing Tumor (ET) and **0.87** for Edema (ED), demonstrating robustness even for heterogeneous tumor morphologies. However, the 2D version adopted to limit computational costs exhibited a common drawback across all models: the loss of volumetric context. Slice-by-slice processing, while efficient in 2D, introduces potential inconsistencies across adjacent slices and fails to fully exploit the anatomical continuity of the tumor, ultimately reducing 3D segmentation accuracy [14].

ResUNet integrates residual connections that enhance gradient flow and training stability in deeper networks [15]. This design is particularly effective for capturing smooth intensity transitions, making it well suited for segmenting necrotic cores and edema regions. In the BraTS19 dataset, ResUNet reached a Dice score of **0.83** for ET and **0.87** for ED. However, a pronounced overfitting was observed for the NCR/NET class (**0.72** Dice in validation versus **0.90** in training), underlining the model's dependence on careful hyperparameter tuning and data quality; without adaptive regularization, the network tends to memorize training patterns rather than generalize.

Attention U-Net introduces spatial attention mechanisms allowing the network to dynamically weight relevant regions during up-sampling, improving sensitivity to small or irregular lesions [16]. In the validation set, the model achieved **0.83** Dice for ET and **0.88** for ED, performing slightly better than ResUNet on edema segmentation. The ability to selectively focus on critical regions is advantageous in complex clinical scenarios, but comes at the cost of significantly increased architectural complexity and training time, without a statistically significant gain over DeepSeg or ResUNet.

Beyond U-Net variants, several **alternative segmentation paradigms** have emerged in the state of the art. The **Segment Anything Model (SAM)** and its medical adaptation **MedSAM** have introduced zero-shot segmentation capabilities to medical imaging, showing promising results in generalizing to unseen modalities [17, 18]. While SAM-based approaches offer flexibility and rapid adaptation, current evaluations on BraTS datasets report Dice scores around **0.80–0.85**, still below dedicated task-specific networks such as nnU-Net. **Vision Transformers (ViT)** and hybrid CNN–Transformer architectures have demonstrated competitive performance in brain tumor segmentation, leveraging self-attention to capture long-range dependencies [19, 20], with reported BraTS Dice scores in the **0.88–0.90** range. Additionally, lightweight detection-oriented models such as **YOLO-based adaptations** have been explored for ROI-based tumor detection in MRI [21], offering speed advantages but generally underperforming in voxel-wise segmentation compared to volumetric CNNs, with Dice scores typically **¡0.80**.

Table 1.1: Comparison of state-of-the-art methods on BraTS datasets (validation Dice scores).

Method	**ET**	**ED**	**NCR/NET**
DeepSeg (2D) [13]	0.87	0.87	0.77
ResUNet (2D) [15]	0.83	0.87	0.72
Attention U-Net (2D) [16]	0.83	0.88	0.72
nnU-Net (3D) [14]	**0.91**	**0.91**	**0.89**
ViT/UNETR [20]	0.89	0.90	0.86
SAM / MedSAM [18]	0.82	0.84	0.80
YOLO-MRI [21]	0.78	0.75	0.70

The comparative analysis highlights several common aspects:

1. **Dependence on manual tuning**: all networks require careful optimization of hyperparameters such as patch size, normalization strategies, and loss functions, making the process dataset-specific and time-consuming.

2. **Intrinsic limitation of 2D approaches**: although computationally efficient, 2D segmentation lacks volumetric coherence, impacting overall 3D performance.

3. **State-of-the-art benchmarks**: dedicated 3D networks such as nnU-Net consistently outperform both U-Net variants and alternative models like SAM and YOLO on BraTS datasets.

These limitations and benchmarks strongly motivate the adoption of **nnU-Net** as the final framework. Unlike manually designed U-Net variants, nnU-Net is a *self-configuring* system that analyzes the dataset properties and automatically adjusts the architecture, hyperparameters, and preprocessing pipeline without manual intervention [14]. This addresses expert-dependent tuning and integrates **3D full-resolution** and **3D cascaded** pipelines, overcoming the main drawbacks of 2D implementations and fully exploiting anatomical continuity. The combination of auto-configuration, optimal use of 3D information, and superior cross-dataset generalization makes nnU-Net a clinically robust choice for multimodal brain tumor segmentation.

1.5 Large language models in healthcare domain

With the advancement of Large Language Models (LLMs), including GPT-4 and its derivatives, a new frontier opens up in the field of explainable artificial intelligence in medicine: conversational explainability. LLMs, in fact, offer the possibility of producing explanations in natural language, structured according to a conversational logic that can come very close to the way an expert doctor explains a diagnosis or a therapeutic choice to a colleague or a patient. This capability can certainly be used in the context of MRI image segmentation for the identification of brain tumours. While the computer vision model can provide a precise output in terms of tumour location and size, the LLM can accompany this output with a comprehensible textual description that contextualises the result, indicates its critical points and answers explicit questions from the physician, such as: 'Why was this area classified as a tumour lesion?' or 'What radiological features led to this segmentation?'. The added value of LLM lies precisely in the possibility of generating flexible responses, adapted to the user's level of knowledge and oriented towards the clinical objective. This approach is consistent with the principles of Human-Centred AI, in which human-computer interaction is not only instrumental, but epistemically significant. Through the integration of LLM into the XAI pipeline, an attempt is made to overcome the limitation of heatmaps, which, although visually showing the areas of activation, do not explain "why" in comprehensible terms for the end user.

In the work that is the subject of this thesis, LLMs were integrated with the aim of using them to generate automatic explanations associated with the results of the segmentation system. This objective, however, requires joint work with the clinical expert and will not be dealt with in this thesis. Instead, in this work we show the results of the LLMs' capabilities, in particular of GPT-4o or OpenAI, to describe MRI images, identify the type of MRI acquisition sent and describe segmentation and heatmap results in comparison. The analysis intends to highlight the actual ability of the model to describe the medical image in a coherent way. It also aims to highlight the need of control and supervision in the use of such systems.

Using LLMs even in a controlled way through prompting does not eliminate the risk of generating erroneous, ambiguous or misleading content. Therefore, their use in clinical settings requires careful initial validation through human supervision mechanisms and iterative feedback from experienced users. Once introduced, it is essential to understand how to use them and how to interpret their signals. Looking forward, the integration of computer vision systems and generative language models represents a promising direction to improve the transparency, acceptability and effectiveness of artificial intelligence systems in healthcare.

1.6 Implementation of the XAI pipeline: from model to GUI

The implementation phase of the pipeline represents the technical core of the performed work, in which I integrated the algorithmic components of segmentation with the explanation, validation and user interaction modules.
The primary objective was to develop an end-to-end system capable not only of accurately identifying brain tumour lesions in MRI images, but also capable of accompanying each decision with an adequate and clinically useful explanation.
The pipeline consists of several phases, each of which contributes to the generation of an interpretable result. The first phase involves the pre-processing of the MRI images. Next, the segmentation algorithm, based on 3D convolutional neural networks (e.g. U-Net, ResNet3D), performs the automatic localisation of the brain tumour. This phase is immediately associated with the XAI module, which uses techniques such as Gradient-weighted Class Activation Mapping (Grad-CAM) [22], Integrated Gradients and TracIn [23] to produce visual and structural explanations of model decisions[2].

[2]In particular, although the TracIn method [24] was studied and demonstrated valuable insights into the influence of individual training data on final predictions—serving as a resource for model validation and clinical risk management—it was not integrated into the main XAI pipeline presented in this thesis. This work is instead reported separately as a complementary study, as discussed in the previous sections.

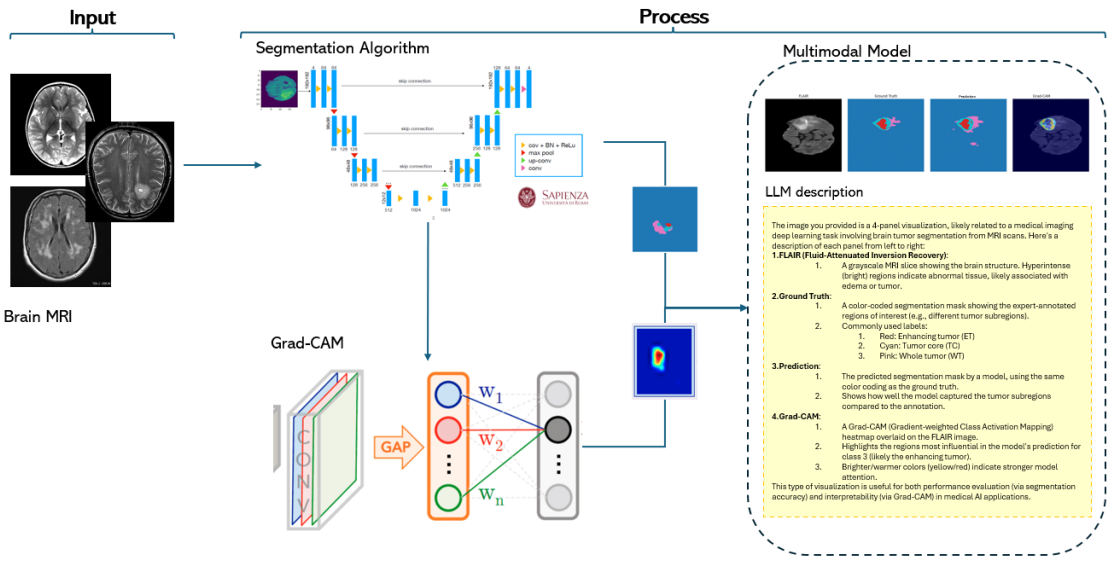


Figure 1.1: Overview of the proposed explainable AI pipeline for brain tumour segmentation in MRI. The pipeline consists of three stages: (1) input preprocessing with multimodal brain MRI sequences; (2) segmentation using a convolutional neural network architecture (U-Net), followed by the application of Grad-CAM to generate class-specific visual explanations; and (3) integration into a multimodal reasoning component powered by a Large Language Model (LLM), which produces human-readable textual descriptions based on the visual inputs. The resulting outputs are presented as a four-panel visualization.

The traceability of decisions makes it possible to trace back similar cases already seen by the model, facilitating the diagnostic process for the physician and increasing confidence in the system. A distinctive aspect of the pipeline is the design of a user-friendly, clinician-oriented graphical user interface (GUI). The interface is organised into three information levels:

- Visualisation of the MRI image with superimposed segmentation.
- Visual explanation generated in real time.
- Textual explanation based on LLM, with the possibility of direct querying by the clinician.

This layered structure allows to navigate from pure numerical evidence to linguistic justification, providing an integrated and modular perspective. the entire structure is modularly designed, so that it can be extended to other pathologies and imaging types, while preserving the scalability and maintainability of the code. The used libraries (PyTorch, MONAI [25], Streamlit) and the adoption of software development practices geared towards reproducible research (model versioning, data tracking, experiment logging) have helped to

ensure the reliability and transparency of the entire architecture.
The proposed pipeline does not produce only computational results, but aims to create a meaningful interaction between artificial intelligence and the clinician, in which explainability is not merely an additional output, but an intrinsic property of the decision-making process.

Chapter 2

Materials and methods

This chapter presents the materials and methodologies used in the development of the explainable AI pipeline for brain tumour segmentation in MRI. The design and implementation of this pipeline followed a modular, user-centred engineering approach, integrating technical, clinical and regulatory considerations.
The chapter is structured to reflect the sequential flow of system components. It begins by describing the datasets used; next, GPT-4o, the multimodal language model used to generate textual explanations from visual data, is presented. The following sections illustrate the experimental procedures developed to evaluate the capabilities of GPT-4o in image interpretation and classification, as well as the integration of segmentation results and Grad-CAM heatmaps into the explanation pipeline. The methodology of the TracIn study is also presented, considered as an explanation technique. Although it was not directly integrated into the pipeline, it was nevertheless included in the XAI studies.
The design and functionality of the graphical user interface (GUI) is then described, with a focus on its role in facilitating clinical validation and interaction. Finally, we present the explainability experiments that evaluate the spatial and semantic coherence of the AI outputs, comparing them against ground truth annotations and examining both visual and language-based justifications.
Together, these components constitute a comprehensive methodology aimed at supporting reliable, interpretable and human-aligned AI in clinical practice.

2.1 Datasets: Br35H and BraTS19

The datasets used in this study are Br35H (Brain tumour Detection 2020) [26], [27] and BraTS19 (Brain tumour Segmentation Challenge) [28].

Br35H consists of 3,000 grayscale brain MRI images, divided equally into two main classes:

- Non-tumourous: images of brains without tumours.
- Tumourous: images containing brain tumours.

The following are the main characteristics of the dataset:

- Image format: JPG.
- Image size: 2D, with a minimum resolution of 176×176 pixels and a maximum resolution of 512×512 pixels.
- Mode: Magnetic resonance imaging.
- Anatomical area: Brain.
- Task type: Binary classification (presence or absence of tumour).

Has been widely used for training and evaluation of deep learning models in the detection and classification of brain tumours.

This subdivision enabled experiments aimed at testing the model's ability to distinguish between healthy and pathological images. In addition, images from the Br35H dataset were used to:

- Test the ability of GPT-4o to recognize MRI: the model was analyzed to determine whether it could correctly identify brain MRI images and distinguish the presence or absence of tumours.
- Evaluating automatic classification without prompt engineering: the results showed that without explicit guidance, the model tends to describe generic abnormalities but does not always accurately distinguish the presence of a tumour.
- Optimization by prompt engineering.

BraTS19 is one of the benchmark competitions in the field of automated brain tumour segmentation on magnetic resonance images (MRI). BraTS is organised annually and provides a benchmark for the development and evaluation of deep learning algorithms dedicated to brain tumour segmentation. The BraTS19 edition focused in particular on the segmentation of gliomas, which represent one of the most common and aggressive neoplasms of the central nervous system [29]. Gliomas are characterised by both structural and biological heterogeneity, with marked intra- and inter-patient variability. This complexity makes essential

the use of advanced methodologies for accurate segmentation of tumour lesions. BraTS19 uses pre-operative MRI images collected from multiple healthcare institutions and clinics, which were manually annotated by certified radiologists. The training dataset consists of 335 patients, divided as follows:

- 259 cases of High-Grade Gliomas (HGG)
- 76 cases of low-grade gliomas (Low-Grade Glioma, LGG)

For each patient, four different MRI modalities are provided, which provide complementary information on the morphology of the tumour and its interaction with the surrounding tissue:

- T1-weighted (T1w) - it highlights the anatomical structure of the brain.
- T2-weighted (T2w) - it enhances visualisation of the presence of oedema and abnormal fluid.
- Post-contrast T1-weighted (T1Gd) - it identifies areas of enhanced tumour.
- T2 Fluid-Attenuated Inversion Recovery (T2-FLAIR) - it visualises areas of peritumoural oedema.

The MRI images have a resolution of 240 $\times$ 240 $\times$ 155 voxels, with 155 slices along the sagittal direction. The manual segmentation provided in the dataset classifies the brain tissue into four categories, each associated with a numerical label:

- Necrosis and non-contrast-enhancing tumour (NCR/NET) - label 1, characterised by a hypointense signal in T1Gd compared to T1w.
- Peritumoral oedema (ED) - label 2, typically evidenced with hyperintense signal in T2-FLAIR.
- Tumour contrast-enhancing (ET) - label 3 (originally label 4 in the BraTS19 dataset), characterised by hyperintense areas in T1Gd compared to T1w.
- Background (BKG) - label 0, representing healthy tissue and non-tumour regions.

Compared to previous editions, BraTS19 has introduced significant improvements in annotation accuracy and image standardisation. Some distinctive aspects of this edition includes:

- An improved segmentation protocol, with greater consistency between human annotators.

- The integration of a larger multi-institutional dataset, which improves the generalisability of the segmentation algorithms.
- The use of advanced metrics to evaluate the performance of deep learning models.

Automated segmentation of brain tumours represents a key step in the clinical management of patients, offering support to radiologists in image analysis and enabling quantitative analysis of disease progression. The presence of multimodality data in the BraTS19 dataset is particularly relevant, as the different MRI modalities provide complementary information on tumour characteristics. In particular, accurate segmentation of different tumour components (NCR/NET, ED, ET) has a direct impact on therapeutic decisions, influencing surgical planning, radiotherapy and monitoring of treatment response. The improvement of automated segmentation methodologies based on deep learning can therefore significantly contribute to the personalisation of therapeutic strategies for patients with gliomas. These datasets served not only as training and validation sources for segmentation models, but also as input material for evaluating the capabilities of multimodal large language models. The next section presents GPT-4o, the LLM used in this study to interpret and generate explanations from these medical images.

2.2 Large Language Model: GPT-4o

The evolution of Large Language Models (LLM) from 2022 to 2024 marked a significant transition from text-only processing to multimodal integration, incorporating both textual and visual inputs. This progression has expanded the capabilities of artificial intelligence models, enabling richer and more versatile understanding and generation of content.
By 2022, LLMs have reached remarkable milestones in natural language processing. Models such as Google's PaLM, with 540 billion parameters, and Meta AI's OPT-175B [30], with 175 billion parameters, demonstrated outstanding performance in translation, summarisation and textual inference tasks. These models focused exclusively on text processing and on improving natural language comprehension and generation.
The year 2023 marked the beginning of multimodal integration in LLMs. OpenAI introduced DALL-E 2 [31] which is capable of generating images from textual descriptions, combining language processing with visual generation. In addition, Google announced Gemini [32], a model designed to be multimodal, capable of processing text, images, audio, video and programming languages simultaneously. This development represented a significant step towards integrating different input modalities into a single artificial intelligence model. A significant next step in this evolution was GPT-4, developed by OpenAI. This model rep-

resents an advancement in the field of Large Language Models (LLMs), employing a new transformer-based architecture with an increased number of parameters, improving language understanding and the generation of more precise and contextually relevant responses.

GPT-4 introduced multimodal capabilities into GPT models for the first time, processing both text and images. The model's training was based on a diverse dataset with fine-tuning techniques for specific tasks, ensuring greater accuracy in visual content analysis. Furthermore, the model was deployed on cloud-based infrastructures with high-performance GPUs, optimising the management of high computational demands. In 2024, OpenAI released GPT-4o, a model that represents a significant evolution from its predecessors, with full multimodal integration. It was designed to simultaneously process text, images, audio and video, improving human-computer interaction through advanced integration of different input modalities. This innovation has enabled a quantum leap in contextual understanding and responsiveness to multimodal requests, moving closer to a holistic understanding of information.
In this thesis, we chose to start with GPT-4 and subsequently GPT-4o. GPT-4o is an advanced language model based on a transformer architecture, designed to improve natural language understanding and text generation. Compared to its predecessors, GPT-4o introduces significant optimisations in parallel processing on GPU and TPU, reducing latency and improving computational efficiency. The architecture is based on a transformer decoder (Figure 2.1) with advanced self-attention mechanisms, allowing more effective context handling on very large text windows, up to 128k tokens.

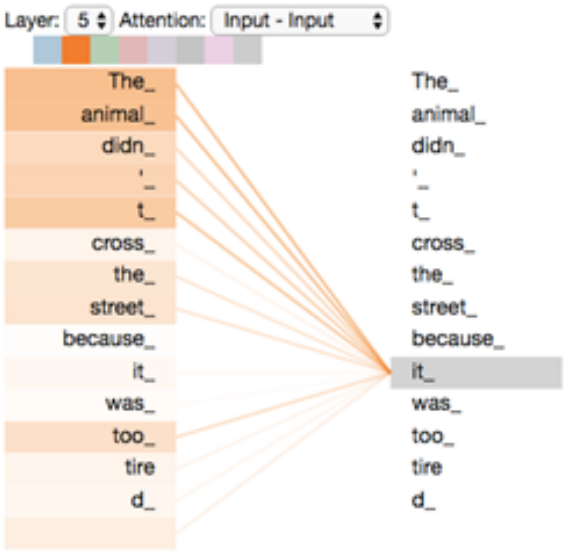


Figure 2.1: Example of self-attention in the decoder of a Transformer model [33]. The token 'it_' (right) assigns different attention weights to the preceding tokens of the sentence. The lines represent the degree of attention devoted to each input term, highlighting how the model is able to capture long-range semantic relationships (e.g. the reference between 'it_' and 'The_ animal_'). This mechanism underlies the model's ability to generate coherent and contextualised text.

Distributed computation plays an essential role in the GPT-4o architecture, exploiting pipeline parallelism and memory optimisation to improve model scalability. This approach enables the model to provide faster and more consistent responses even on highly complex texts. Another innovative feature of GPT-4o is its multimodal processing capability [34], that allows it to interpret and generate textual content based on visual inputs, such as images and diagrams. This functionality makes it particularly useful in the medical field, where it can support professionals in diagnosing and explaining MRI images.
GPT-4o's architecture also includes an improvement in natural language generation, thanks to a more advanced handling of text consistency and text stylistics. This allows the model to adapt to different communicative contexts, ensuring greater accuracy in responses. The model was trained on a large dataset, including a variety of technical and scientific texts, in order to improve its semantic and contextual comprehension capability. Furthermore, the implementation of training optimisation strategies, such as the use of distributed parallelism and the dynamic adjustment of learning parameters, led to a more robust model which is able to adapt quickly to new data and domains.
GPT-4o represents an important development in the field of artificial intelligence applied to natural language processing.

Thanks to its optimised architecture and multimodal capabilities, the model lends itself to multiple applications, from academic and medical support to the generation of specialised content.
The following section describes the experimental procedures developed to assess how this model performs in interpreting and classifying medical images from the selected datasets.

2.3 Experimental Analyses with GPT-4o and Clinical Data

In this section, the main experiments conducted to evaluate the capabilities of the multimodal GPT-4o model in the automatic classification of magnetic resonance images are presented and analysed. The experiments focused on two distinct areas: the automatic recognition of tumour images from the BR35H dataset and the classification of MRI acquisition modalities using the BraTS19 dataset.

These experiments were conducted with a zero-shot prompt and, based on the results obtained, the prompt was modified to obtain the best performance or to highlight any criticalities encountered.

Various tools and technologies were used to perform these analyses, including:

- datasets **BR35H** and **BraTS19**, used for model training and validation;
- the OpenAI multimodal language model **GPT-4o** [35];
- languages and development environments such as **Python** and **Visual Studio Code**;
- a non-relational database management system **MongoDB** for storing and analysing results. The choice of MongoDB as a management system is motivated by its flexible and schemeless nature, ideal for containing semi-structured data with base64-encoded images and textual metadata. In addition, its document structure allows rapid indexing and querying of results associated with specific image id, modality, or channel, facilitating aggregated analysis.

The following paragraphs describe in detail the procedures adopted.

2.3.1 Automatic classification of tumour images in the BR35H dataset

To initiate the evaluation, a zero-shot experiment was conducted to test the multimodal capabilities of GPT-4o in distinguishing brain MRI images with tumour lesions from those without. In artificial intelligence, zero-shot learning refers to a model's ability to perform tasks it has not been explicitly trained for, by leveraging knowledge acquired in different contexts [36]. This is particularly relevant for large language models, which can interpret and respond to clinical prompts even without domain-specific fine-tuning. After the zero-shot analysis, different prompts where tried, based on the obtained results.

Phase 1 - Recognition of tumour images by GPT-4o

The experiment focused on assessing whether GPT-4o could correctly identify the presence or absence of tumours in brain MRI images from the BR35H dataset, which includes images labeled as either 'tumour' or 'no-tumour'. A series of five sub-experiments were designed to evaluate the model's classification abilities under different prompt configurations.
Initially, separate analyses were carried out for the two classes. In the first sub-experiment, all 'no-tumour' images were submitted to the GPT-4o API using a simple descriptive prompt:

Prompt: "Describe the image"

The goal was to observe whether the model, without explicit instruction, would tend to report abnormalities in images known to be healthy. In the second sub-experiment, the same prompt was applied to the 'tumour' class, to assess the model's natural ability to detect and describe pathological features.
To enable a more standardised evaluation, a third and fourth sub-experiment were conducted using a binary classification-oriented prompt:

Prompt :'Describe the image and indicate whether a brain tumour is present or not.'

This prompt was applied separately to the 'no-tumour' and 'tumour' images to allow direct comparison between model predictions and ground truth.
Finally, in the fifth and most comprehensive sub-experiment, the full dataset was analysed in random order using the binary prompt, simulating a more realistic diagnostic scenario. Each image was processed once; all descriptions were then automatically analysed by GPT-4o and labelled with an additional prompt: "Here is the description of a brain MRI image: answer strictly with one of the following options: 'tumour', 'no tumour' or 'unknown'" in order to save a direct label to compare with ground truth labels and subsequently compute accuracy,

precision, recall, F1-score, and confusion matrices.

This multi-step approach provided a quantitative assessment of the zero-shot classification performance of GPT-4o on non-standardised clinical images and allows one to realise how much the slightest change in the prompt can change the response completely or not, highlighting how delicate and important this practice is.

Phase 2 - Automatic detection with bounding box

In parallel, a second analysis model was integrated, specifically trained for the automatic detection of brain tumours, available on the Roboflow platform (model 'brain-tumour-ovfqd/1') [37]. The object detection model was applied to each image in the BR35H dataset via a Python interface based on 'inference sdk' [38]. The result was the automatic generation of bounding boxes around suspicious areas (Figure 2.2), associated with a classification probability.

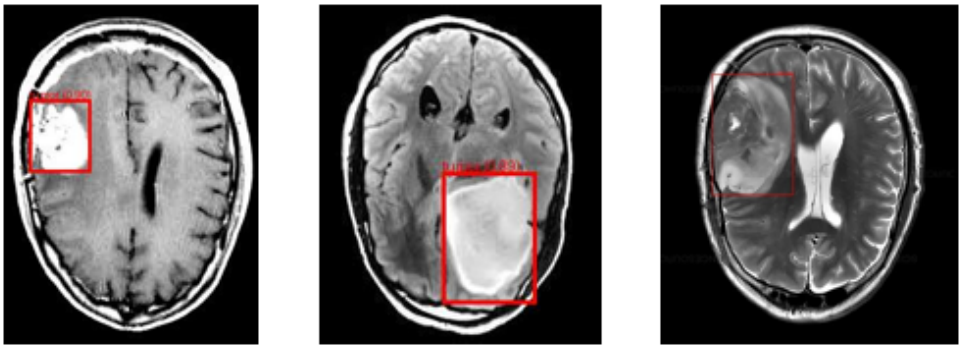


Figure 2.2: Generation of boundings boxes around suspicious areas (images from BR35H dataset).

The annotated images were saved locally in a structured directory, while the related metadata (file name, annotated image path, id, class detected, confidence) was recorded in a structured JSON file (Appendix A), useful for subsequent validation and visualisation in the graphical interface. The model's error rate was also calculated, comparing the results obtained with the original labels of the dataset and applying classification-type metrics.

Phase 3 – Quadrant-based localisation of tumour areas

A further experimental phase was developed to explore whether GPT-4o, beyond simple binary classification, could also support the spatial localisation of tumour regions within brain MRI images. This task aimed to assess the model's interpretative capacity in a more structured visual context, by asking it not only to detect the presence of a tumour, but also to indicate in which specific area of the image it was located. For this purpose, all the images labelled as "tumour" in the BR35H dataset were selected and pre-processed to include a visual subdivision into four quadrants. Each image was overlaid with red horizontal and vertical lines intersecting at the centre, effectively dividing it into four equal sections, numbered from 1 to 4 as shown in the Figure 2.3.

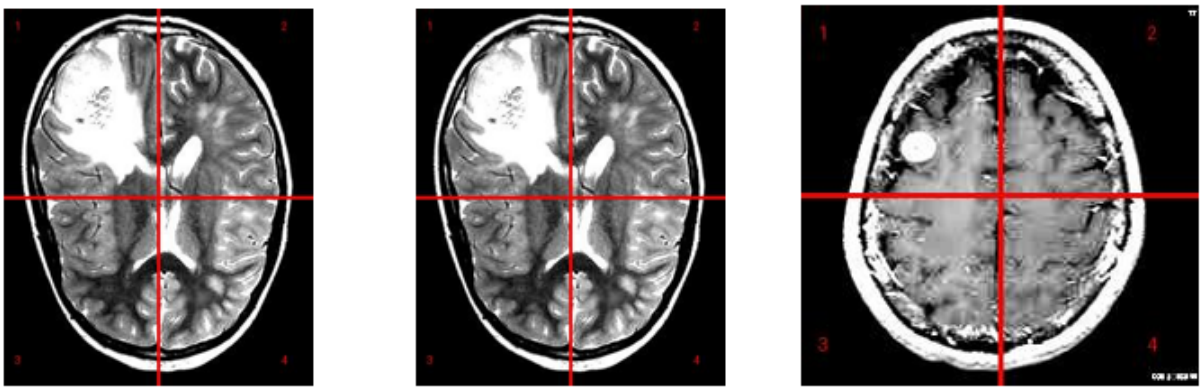


Figure 2.3: Images from BR35H dataset divided in 4 enumerated quadrants

This visual structure was intended to offer a simplified spatial framework that GPT-4o could potentially interpret, even without prior training on spatial annotation tasks.
A targeted prompt was used when submitting the annotated images to the model: the request explicitly asked for a binary diagnosis (presence or absence of tumour) and, for in positive cases, the indication of the quadrant or quadrants involved. The model's responses were parsed to extract both the diagnostic judgement and the localisation information.
The entire process — from image preparation and visual annotation to model interaction and response recording — was implemented via a custom Python script, which also managed the automatic storage of predictions in a dedicated MongoDB collection. Details on the full pipeline and implementation can be found in the Appendix .
To ensure the reliability of the results, all model responses were manually reviewed. A second script was developed to guide the researcher through the visual inspection and validation process. For each image, the original GPT-4o response was compared with the human judgment regarding both the presence of the tumour and the accuracy of quadrant selection.

This step was essential to construct a validated reference dataset, enabling the calculation of detailed performance metrics.
This experiment introduced a higher level of interpretative complexity and was designed to test the potential of multimodal language models in approximating not only diagnostic, but also spatial reasoning in medical imagery. The quantitative results of this analysis, including quadrant-level error rates and agreement with human validation, are discussed in the Results section.

2.3.2 Semantic Evaluation of Descriptions

To evaluate the textual outputs generated by GPT-4o, different semantic evaluation strategies were applied depending on the phase of the experiment. In paragraph 2.3.1 and paragraph 2.3.1, GPT-4o's free-text responses were semantically interpreted and automatically assigned a binary label (*tumour* or *no tumour*). This labeling process enabled the application of standard quantitative evaluation metrics, such as accuracy, precision, and recall, as detailed in Section 2.5. As regards the paragraph 2.3.1, which involved quadrant-based localisation of tumour areas, required a more targeted, spatially grounded evaluation. In this case, the assessment focused on verifying whether the quadrants identified by the model corresponded to the actual tumour regions. Each response was manually reviewed and evaluated by answering a simple binary question: "Are the indicated quadrants correct?" This approach allowed for flexible yet consistent interpretation in the absence of structured labels or segmentation masks.

2.3.3 Automatic classification of MRI modalities by GPT-4o

An analysis was also conducted to assess GPT-4o's ability to correctly classify brain magnetic resonance imaging (MRI) acquisition modes, operating in zero-shot mode and without the aid of contextual information. The experiment involved presenting the model with images from the four sequences in the BraTS19 dataset - T1w, T2, T1gd (T1-weighted with contrast medium) and T2-FLAIR - to test whether it was able to distinguish among them solely by visual information.
This type of evaluation is particularly relevant for exploring the applicability of multimodal models in unsupervised clinical scenarios, where the accuracy of visual interpretation may directly affect the reliability of the system in decision support.
The following procedure was adopted to carry out the experiment:

1. Image retrieval: the 96 validation set samples from the BraTS dataset were retrieved from the MongoDB decathlon_dataset collection, each containing 4 slices, one for each MRI modality (FLAIR, T1w, T1gd, and T2). Channel mapping: Translation of the channel_i key into the corresponding MRI modality (Appendix A).

2. Sending to GPT-4o: converting the slice to base64 format and sending to the GPT-4o API with a specific prompt.

3. Parsing of the response: Analysis of the text returned by GPT and normalization of the prediction.

4. Saving the results: Inserting the data into a new MongoDB modality_predictions collection.

Ground truth is then inferred directly from the key associated with each image slice. In fact, during pre-processing, each multimodal channel is separated into 2D slices and encoded in base64, keeping the channel position as the key identifying the acquisition type. Since the position is fixed and known within the BRATS19 dataset, it is possible to associate each slice with its original MRI modality.

A summary table is given below (Table 2.1):

Table 2.1: Correspondence between channel keys (channel_i) and MRI acquisition mode in the BraTS19 dataset

Key channel_i	Associated MRI mode	Summary description
channel_0	FLAIR	Highlights peritumoral edema by suppressing CSF
channel_1	T1w	Shows the anatomical structure of the brain
channel_2	T1gd	Highlights the tumour at the contrastographic stage
channel_3	T2	Makes visible the presence of fluid and areas of oedema

To manage the results of the experiment, a new MongoDB collection called modality_predictions was created. It is constructed within the predict_and_store() function of the Python script, which reads each document from the decathlon_dataset source collection, extracts the images associated with each channel, decodes them and sends them to the GPT-4o model via the API.

Each image is converted to base64 format and attached to a request that includes the following prompt:

Prompt = "Based only on the image, identify the MRI modality type. Choose from: FLAIR, T1w, T1gd, or T2."

The model returns a textual description, from which the prediction is extracted using a keyword-based filter. The prediction is then compared with the ground truth derived from the channel_i key and saved as a structured document in the modality_predictions collection (Appendix A).
The code also provides for the collection of all predictions in a results list, which is subsequently analysed by the analyse_results() function. At this stage, a pandas Data Frame is constructed, exported in CSV format, and the accuracy metrics and confusion matrix are calculated.

Finally, a graphic visualisation of the results is generated via a heatmap of the confusion matrix and saved as a .png image. This process allows a quantitative evaluation of the behaviour of the GPT-4o model in the visual recognition of MRI modalities. The choice of MongoDB as a management system is motivated by its flexible and schemeless nature, ideal for containing semi-structured data with base64-encoded images and textual metadata. In addition, its document structure allows rapid indexing and querying of results associated with specific image id, modality, or channel, facilitating aggregated analysis.

2.4 Explainability Experiments

In addition to the experiments aimed at investigating the potential and limitations of GPT-4o in generating accurate and consistent descriptions in the relevant clinical context, further experimental tests were developed to investigate different explainability methodologies.
One of the main experiments focuses on evaluating the spatial coherence between activation maps generated by Grad-CAM and the reference segmentation masks in the BraTS19 dataset. The aim is to verify to what extent the regions identified as relevant by the classifier coincide with the actual tumour areas, thus providing a quantitative measure of the reliability of the visual explanations produced by the model.
A second experiment instead explores the use of TracIn [24] as an alternative approach for explainability, with the aim of assessing its effectiveness in tracking the influence of training data on the model's final predictions.

2.4.1 Heatmap vs. Segmentation Correspondence

To evaluate the interpretability of segmentation predictions, the Grad-CAM (Gradient-weighted Class Activation Mapping) technique is adapted to a 2D convolutional neural network architecture based on MONAI. Since the original segmentation model processes volumetric (3D) input, each MRI volume from the BraTS19 dataset is decomposed into axial 2D slices. These slices are normalized and resized as necessary to match the input format of a MONAI-based 2D U-Net model.

The network accepts four-channel input, corresponding to the T1, T1ce, T2, and FLAIR MRI sequences, and performs multiclass semantic segmentation, predicting one of four anatomical classes for each pixel: background (BG), necrotic core (NCR), edema (ED), and enhancing tumour (ET).

Grad-CAM is applied to the final convolutional layer of the U-Net decoder to produce class-specific heatmaps. The method highlights the regions of the input image that contribute most to the prediction for a chosen target class (e.g., enhancing tumour). The heatmap is computed by weighting the feature maps according to the gradients of the target class score, using the formulation:

$$L^{c}_{\text{Grad-CAM}} = \text{ReLU}\left(\sum_{k} \alpha_k A^k\right)$$

where A^k is the k-th activation map and α_k is the global average of the gradient of the class score y^c with respect to A^k.

To assess the spatial correspondence between the model's attention and the ground truth, the generated Grad-CAM heatmap is binarized and compared to the segmentation mask for the same target class. Two standard overlap metrics are used: *Intersection over Union (IoU)* and *Dice Similarity Coefficient (DSC)*, defined as:

$$\text{IoU} = \frac{|H \cap S|}{|H \cup S|}, \quad \text{DSC} = \frac{2 \cdot |H \cap S|}{|H| + |S|}$$

where H denotes the binarized heatmap and S the ground-truth segmentation mask for the target class.

For quantitative evaluation, the Grad-CAM heatmaps were binarized before computing the IoU and Dice scores. Following standard practice for gradient-weighted class activation maps in medical imaging [39, 40], all heatmaps were first normalized to the [0,1] range on a per-image basis and then thresholded at a fixed cutoff of 0.5. The same threshold was applied across all experiments and segmentation classes (background, edema, enhancing tumor, NCR/NET), and no class-specific or dataset-specific tuning was performed to maintain

reproducibility. This approach provides a structured and quantitative framework to evaluate whether the model's explanations focus on clinically meaningful regions, supporting transparency and trust in the AI-assisted diagnostic workflow.

2.4.2 Influence-based Explainability for Segmentation

Following the preliminary state-of-the-art analysis of the performance of segmentation models, a targeted selection of the most suitable algorithm for this research was carried out. While the integration of an explanation method based on TracIn [24] offers interpretative potential by quantifying the influence of individual training examples on model predictions, it also introduces significant computational complexity.

TracIn was investigated as part of a separate exploratory study and was not integrated into the final XAI pipeline evaluated in this thesis.The analysis contributes to a broader understanding of explainability in medical image segmentation.

To balance interpretability with computational efficiency, three advanced segmentation models—DeepSeg, ResUNet, and Attention UNet—were implemented in their two-dimensional (2D) configurations.

The 2D implementation treats each axial MRI slice independently. Multimodal MRI acquisitions (T1, T1Gd, T2, T2-FLAIR) are used as input channels, analogous to RGB image channels, allowing the network to exploit the complementary information across modalities. Each pixel is assigned to one of four mutually exclusive classes: 0 (background - BKG), 1 (non-contrast-enhancing/necrotic tumor - NCR/NET), 2 (edema - ED), 3 (contrast-enhancing tumor - ET), producing a predicted segmentation mask for each class.

To reduce computational load and information redundancy, only 10 central axial slices were selected per patient. This choice was validated empirically and aligns with findings from recent work [23], showing that this subset captures sufficient tumor volume while keeping explainability methods tractable. The final training dataset included 2070 examples, with 520 held out for testing.

Additional strategies were employed to enhance robustness: images were resized from 240×240 to 192×192 pixels; intensity values normalized in the range [-1, 1]; and elastic data augmentation (random cropping and mirroring at 50% probability) applied to improve generalization. The softmax function was used to normalize the network's output logits.

Performance evaluation relied on the Dice coefficient, a metric well-suited to medical segmentation where class imbalance (e.g., small tumor vs. large background) is common. The loss function excluded the background class and was computed as the average Dice loss over the tumor classes:

$$L = 1 - \frac{1}{3} \sum_{i \in Cl} D_i \tag{2.1}$$

where Cl represents the set of segmentation classes (in this case $Cl = 1, 2, 3$) and D_i is the Dice score for class i.

An extension of the TracIn algorithm was implemented for segmentation tasks, originally designed for classification. The adapted method computes class-specific influence scores, identifying the most impactful training examples (proponents/opponents) for each predicted region. To improve relevance, the influence computation was restricted to pixels with softmax confidence above 0.8. This adaptation allows for a clearer association between predicted segmentations and their supporting training examples, addressing the opacity of deep learning decisions in clinical applications.
Originally proposed for classification tasks, TracIn (Training Influence) estimates the impact of individual training samples on a model's predictions by accumulating gradient similarities over the course of training. For a training example z, a test example z_{test}, and model parameters $\theta^{(t)}$ at training step t, the influence is computed as:

$$\text{TracIn}(z, z_{\text{test}}) = \sum_{t \in T_z} \eta^{(t)} \cdot \nabla_\theta \mathcal{L}(z; \theta^{(t)})^\top \cdot \nabla_\theta \mathcal{L}(z_{\text{test}}; \theta^{(t)}) \tag{2.2}$$

where $\eta^{(t)}$ is the learning rate at step t, and T_z is the set of training steps in which the training example z is used. This formulation captures the alignment of the gradient directions, indicating whether a training example contributes positively (proponent) or negatively (opponent) to a given prediction.
To enable practical computation on large datasets, gradient evaluations are performed at a limited number of saved training checkpoints, bypassing the need for second-order derivatives. This approximation, while efficient, preserves the interpretability of the influence scores and has been shown to align with human intuition in identifying mislabeled or highly influential samples. The adaptation to segmentation contexts thus extends TracIn's utility to structured prediction tasks, offering a mechanism for tracking evidence attribution at the pixel or region level.

Although the adaptation of TracIn for segmentation demonstrated promising potential in identifying influential training examples, it was ultimately not integrated into the final XAI pipeline evaluated in this thesis.Consequently, no results are reported regarding its potential combination with Large Language Models (LLMs).

Integration of TracIn into the XAI Pipeline. In addition to the Grad-CAM based visual explanation module, the pipeline design also incorporates the TracIn influence method as an alternative XAI component [41]. Unlike feature attribution approaches such as Grad-CAM, TracIn provides a data-centric form of explainability by identifying the most influential training examples (proponents and opponents) that contributed to the prediction of each segmented region.

Within the architecture of the explainable AI pipeline (Figure 2.4), this integration introduces a dedicated branch linking the segmentation model to the training data repository for the computation of influence scores on demand. The resulting ranked examples are processed by a Large Language Model (LLM) to generate human-readable explanations, connecting the prediction to similar cases encountered during training and presenting the rationale in a form closely aligned with clinical reasoning [42].

On the GUI side, TracIn is implemented as an optional tab complementing the Grad-CAM visualisations: for any selected tumor region, the user can access the list of top influential cases together with the LLM-generated explanation. This addition requires a corresponding update to the validation workflow: the questions posed to the clinical validator are adapted to evaluate the new data-centric explanation paradigm, focusing on the consistency between influential cases and the predicted segmentation.

By combining TracIn with an LLM-generated narrative layer, the pipeline extends beyond activation-based saliency to provide a hybrid explanation that is not only quantitatively grounded but also cognitively aligned with the way human experts justify diagnostic decisions.

2.4.3 Language-based Description Evaluation

Once visualisations are generated, they are provided as input to GPT-4o, initially without any explicit prompts and later with prompts added, in order to assess the model's ability to generate coherent descriptions in the absence of instructions, as well as to highlight its sensitivity to prompt variation. The experiment includes both simple and structured prompts (as shown in Table 2.2), with the aim of investigating the model's adaptability to a highly specific and complex task, such as the visual interpretation of medical images.

In evaluating the textual outputs generated by GPT-4o, we employed a manual assessment approach due to the complex of this task. This involved a structured rubric focusing on relevance, clarity, completeness, and consistency, aligning with established frameworks for LLM evaluation in healthcare. This manual evaluation approach is consistent with current

Table 2.2: Prompt variants used for language-based description generation

ID	Prompt Text
1	*Describe the image.*
2	*Describe the image without assuming any labels or prior knowledge.*
3	*Describe the image. It is composed of four panels arranged from left to right:* *An MRI slice from the BraTS19 dataset. The modality is one of the following: FLAIR, T1, T1Gd, or T2.* *The corresponding ground truth segmentation mask from the BraTS19 dataset.* *The prediction output from a MONAI-based 2D segmentation model.* *A Grad-CAM heatmap generated using the same model.* *Focus your description strictly on the visual content of the image. Do not make assumptions beyond what is explicitly shown. The context relates to MRI analysis of the brain, potentially involving tumor presence, healthy tissue, or other neurological conditions.*

best practices in assessing LLM performance in medical contexts [43, 44].
We defined a structured set of criteria focused on core aspects of the generated descriptions. Each description was assessed along five dimensions:

- **Relevance**: Does the description accurately reflect the contents of the image?
- **Accuracy**: Does the description stay grounded in the visible content of the image, without introducing information that cannot be visually confirmed?
- **Clarity**: Is the text clearly phrased, grammatically correct, and unambiguous?
- **Completeness**: Does the description mention all relevant elements in the image, including each panel?
- **Consistency**: Is the information internally coherent and logically structured?

Each criterion was scored using a four-point Likert-style scale [45] to reflect the quality and adequacy of the response:

The results generated by GPT-4o, the segmentation models and the heatmap are not intended to be used independently. For clinical integration, a structured graphical user interface was developed to allow clinicians to view these results, interact with the segmentation results, become familiar with the use of the heatmap and access the textual descriptions. Section 2.6, presents the design and structure of this user interface.

Table 2.3: Manual Evaluation Scoring Scale (Likert-style)

Score	Interpretation
1	Poor / Incorrect
2	Partially correct or incomplete
3	Mostly correct and clear
4	Excellent / fully correct, clear, and complete

2.5 Evaluation Metrics

To assess the performance of the components of the explainable AI pipeline, we adopt standard evaluation metrics tailored to the nature of each task. The following metrics are used depending on whether the task involves classification, localisation, or spatial segmentation.

Accuracy – Proportion of correct predictions over the total number of predictions:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{2.3}$$

Per-class Accuracy – Proportion of correct predictions computed individually for each class:

$$\text{Per-class Accuracy}_i = \frac{TP_i}{TP_i + FP_i + FN_i} \tag{2.4}$$

Precision – Proportion of true positives among all predicted positives:

$$\text{Precision} = \frac{TP}{TP + FP} \tag{2.5}$$

Recall (Sensitivity, TPR) – Proportion of true positives among all actual positive cases:

$$\text{Recall} = \frac{TP}{TP + FN} \tag{2.6}$$

F1-score – Harmonic mean of precision and recall, useful for handling class imbalance:

$$\text{F1-score} = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{2.7}$$

True Negative Rate (TNR) – Proportion of correctly predicted negatives among all actual negative cases:

$$\text{TNR} = \frac{TN}{TN + FP} \tag{2.8}$$

False Positive Rate (FPR) – Proportion of false positives among all actual negatives:

$$\text{FPR} = \frac{FP}{FP + TN} \tag{2.9}$$

False Negative Rate (FNR) – Proportion of false negatives among all actual positives:

$$\text{FNR} = \frac{FN}{FN + TP} \tag{2.10}$$

Confusion Matrix – A matrix representation of model predictions versus true labels, showing counts for each true/predicted class pair. It is particularly useful in multi-class classification to visualise specific confusions between classes.

Intersection over Union (IoU) – Spatial overlap between predicted and ground-truth regions:

$$\text{IoU} = \frac{|H \cap S|}{|H \cup S|} \tag{2.11}$$

Dice Similarity Coefficient (DSC) – A spatial overlap metric commonly used in medical image segmentation:

$$\text{DSC} = \frac{2 \cdot |H \cap S|}{|H| + |S|} \tag{2.12}$$

These metrics are applied consistently across the experimental results presented in Chapter 3, with specific combinations chosen according to the task type.

2.6 User Interface

The implementation of a user interface (UI) is an essential element for the validation and clinical integration of Explainable Artificial Intelligence algorithms in MRI image segmentation for brain tumour detection. The design of the UI was based on the fundamental principles of XAI, namely transparency, accountability, privacy, completeness of information, and appropriateness to the end user's level of technological knowledge [10], [46].
A well-designed interface not only facilitates interaction with AI models, but also allows for the collection of medical feedback to improve the robustness and reliability of the segmentations produced by the model [1], [47].
The interface was developed in Python using the PyQt5 framework for building the GUI, with Visual Studio Code as the development environment. Representative screenshots of the GUI and a detailed description of the workflow are provided in Appendix E to support the discussion on the originality and contribution of the developed interface. It is organized into four main sections, each designed to support specific steps in the clinical segmentation and validation process:

1. Documentation: it provides detailed information on the dataset used, the XAI algorithms adopted and the evaluation metrics, ensuring transparency in the use of the system. This section contains scientific articles and reference materials to support medical decision making [48].

2. Dataset analysis: it allows users to add new images for validation, improving the ability of the AI model in recognizing morphological changes in brain tumours. It also allows the user to examine statistics related to the datasets used and compare the results of AI segmentations with those of manual radiological annotations [29].

3. XAI Tool: the heart of the system, this section includes the interactive interface for visualizing the segmentations produced by the algorithm. Here, clinicians can explore the results, analyze the textual explanations provided by the XAI system, and evaluate the level of reliability of the predictions. Users can approve, correct or reject the segmentations, thus contributing to continuous improvement of the model.

4. Clinical support tools: it consists in manual segmentation software, such as ITK-SNAP which is widely adopted by radiologists to correct AI-generated masks. This functionality is crucial to ensure that the AI model can be validated directly by experts, increasing its acceptance in the clinical setting [49].

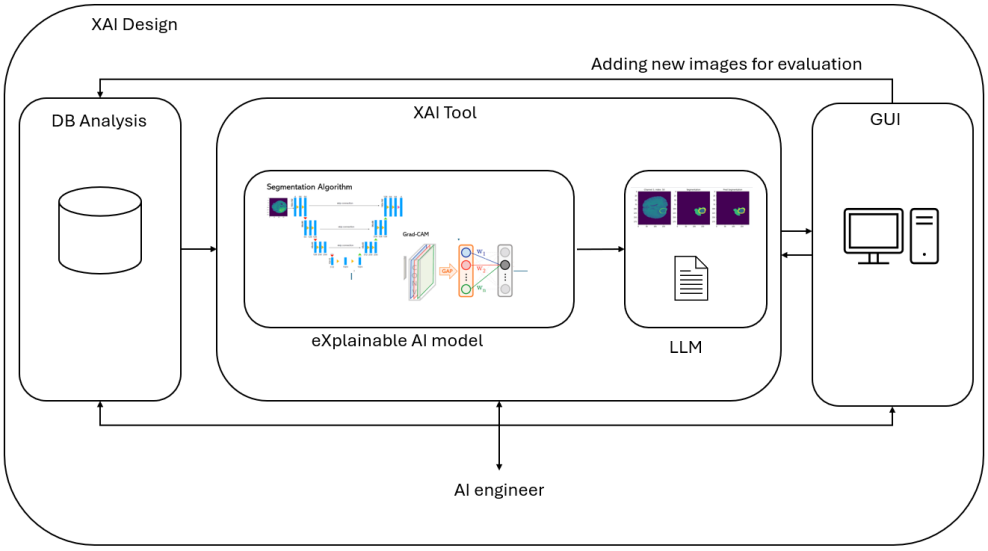


Figure 2.4: System architecture of the explainable AI pipeline for brain tumor segmentation in multimodal MRI. The framework is structured within a top-level block labeled "XAI Design" and includes four main components: (1) DB Analysis, which manages data extraction and pre-processing; (2) XAI Tool, where a UNet-based segmentation model processes MRI inputs and is coupled with explainability methods; (3) an LLM module that generates textual outputs or justifications based on the model's results; and (4) a GUI that allows expert users to visualize, validate, and interact with the segmentations. The loop enables iterative model refinement with user feedback and new image input.

One of the most significant innovations of the UI is the integration with Large Language Models (LLMs), specifically GPT-4o, to transform AI segmentations into textual descriptions understandable by clinicians. The goal of this functionality is to reduce the interpretive barrier between the AI model and the clinician, allowing a smoother and more immediate interaction with the XAI-based decision support system [35]. Use of the interface follows a structured workflow:

1. Selection of the MRI image to be analysed.
2. Generation of the AI segmentation with visualization of the result.
3. Evaluation of the segmentation by the physician, with the possibility to confirm, modify or reject the result.
4. Generation of textual explanation of segmentation using GPT-4o.
5. Interaction with manual correction tools such as ITK-SNAP to validate the results. This pipeline allows the AI model evaluate the AI model, the XAI model (or models) used and the potential integration into the radiology workflow without altering

established practice, ensuring that results are always verifiable and understandable [50].

Chapter 3

Results and Discussions

This chapter presents the experimental results obtained from the implementation of the explainable AI pipeline described in Chapter 2. While the system has been developed taking into account clinical applicability, it has not yet been evaluated by clinical users.
The results presented focus on the technical performance of individual components and their combined potential for supporting interpretability in brain tumour imaging.
The first part of the chapter reports the outcomes of the experimental evaluation of GPT-4o in the generation of textual explanations and image classification. Multiple scenarios were explored, including zero-shot prompting, prompt-guided tasks, and variations in the visual input format (e.g., FLAIR images alone, segmentation overlays, and Grad-CAM heatmaps). A comprehensive evaluation was conducted using standard classification metrics—accuracy, precision, recall, F1-score, and confusion matrices—to quantitatively assess model behaviour under different prompting scenarios. Full definitions and reporting are provided in the subsequent sections.
Subsequently, we evaluate the correspondence between Grad-CAM heatmaps and segmentation outputs, focusing on their spatial alignment using standard overlap metrics such as the Dice Similarity Coefficient (DSC) and Intersection over Union (IoU). This analysis aims to determine whether the model's attention[1] is concentrated on clinically relevant regions.
Finally, the graphical user interface (GUI) is described from a functional perspective.While the interface integrates all core components of the pipeline and supports user interaction, editing, and feedback, its evaluation is currently limited to internal testing and visual demonstration, without formal clinical validation.

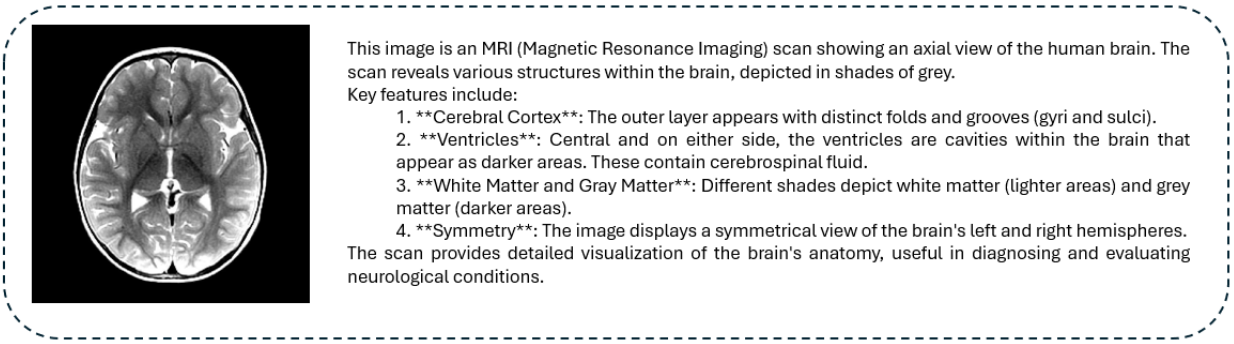
[1]The term "attention" is used here to describe the regions highlighted by Grad-CAM, which estimates the importance of image regions based on gradient information flowing into the final convolutional layers. This should not be confused with the "attention mechanism" found in transformer architectures, which compute weighted feature representations through self-attention. Grad-CAM is not based on transformers and does not involve self-attention or token-based structures.

3.1 Recognition of tumour images by GPT-4o

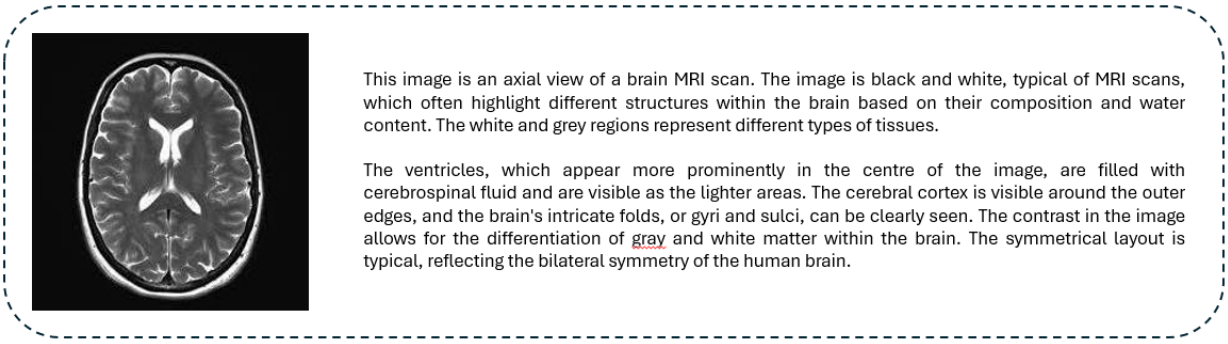
In the first experiment conducted on the no-tumour class of the BR35H dataset, the behaviour of the multimodal GPT-4o model in the absence of specific classification instructions was evaluated using the generic prompt "Describe the image". The aim was to check whether the model, although not explicitly guided, could misinterpret images without lesions as pathological.

A total of 1,497 images belonging to the no-tumour class were analysed. Of these, GPT-4o correctly identified the absence of tumour in 1,376 cases, resulting in True Negatives (TN), while in 121 cases it incorrectly attributed the presence of a tumour, thus generating False Positives (FP).

The Figure 3.1 shows two examples of descriptions of two MRI labeled "no-tumour".



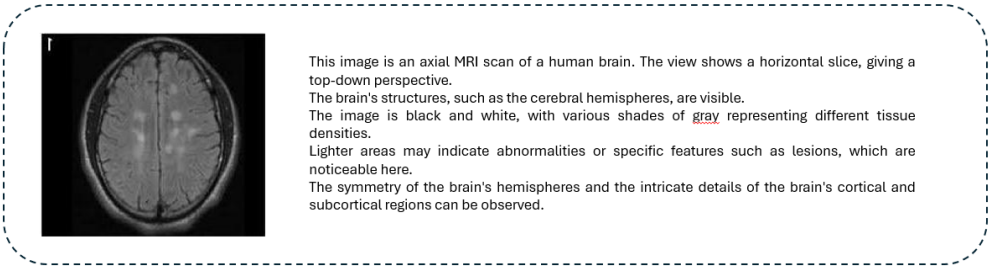
(a)



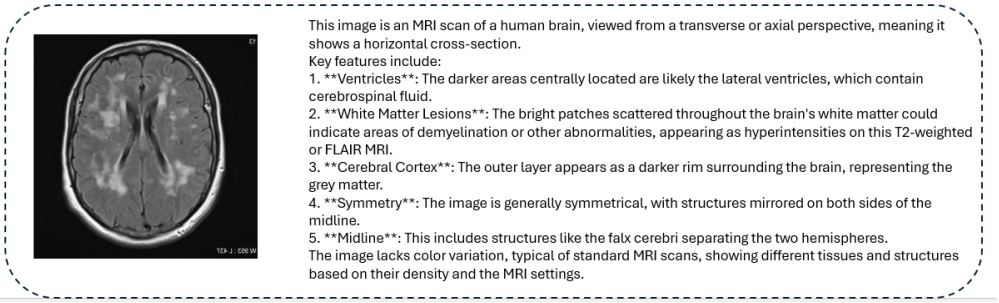
(b)

Figure 3.1: Example of automatically generated descriptions by gpt-4o of class "no-tumour" with prompt 'describe the image', labeled correctly "no-tumour".

The following descriptions, on the other hand, show how GPT-4o does not correctly interpret when lesions or abnormalities are present within the generated description (Figure 3.2).



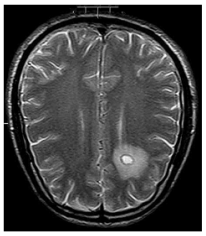
(a)



(b)

Figure 3.2: Example of automatically generated descriptions by gpt-4o of class "tumour" with prompt 'describe the image', labeled wrongly "tumour".

The experiment was also extended to the tumour class of the BR35H dataset, with the aim of testing the ability of the GPT-4o model to correctly recognise the presence of tumour masses in MRI images, again with the generic prompt "Describe the image". 1,485 images containing brain tumours were analysed (Figure 3.3 and 3.4).



This image appears to be an MRI scan of a human brain, shown in a transverse (axial) section. It displays distinct features of the brain, including the cerebral cortex and the white matter. In the lower right area, there is a noticeable bright spot, which might indicate an abnormality such as a lesion or tumor. The greyscale of the image is typical for MRI scans, providing detailed visualization of the brain's internal structures.

(a)

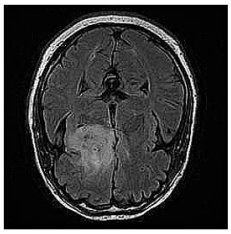
The image is an MRI scan of the human brain. It shows a detailed cross-sectional view. Notably, there is an abnormality present in the upper right area of the brain, which appears as a lighter region compared to the surrounding tissue. This could indicate a potential issue such as a tumor or lesion. The rest of the brain structures, including the folds and lines representing the gray and white matter, are visible throughout the scan..

(b)

Figure 3.3: Example of automatically generated descriptions by gpt-4o of class "tumour" with prompt 'describe the image', labeled "tumour".

The image is an MRI scan of the human brain, presented in an axial view. It shows the brain's internal structures, including the cerebral hemispheres, ventricles, and sulci. The scan highlights different tissues and areas, with varying shades of gray indicating different densities.

(a)



The image is a brain MRI scan showing a transverse section. It depicts various regions of the brain, including the cerebral hemispheres and lateral ventricles. The scan highlights differences in tissue density, which may indicate the presence of certain conditions or abnormalities.

(b)

Figure 3.4: Example of automatically generated descriptions by gpt-4o of class "tumour" with prompt 'describe the image', labeled "no-tumour" (FN).

The model correctly identified the presence of the lesion in 1,349 cases, corresponding to True Positives (TP), while it failed in 136 cases, misclassifying them as normal (False Negatives, FN). Overall accuracy was computed by treating "unknown" responses[2] as incorrect predictions.

False Positive Rate (FPR) and False Negative Rate (FNR) were instead computed *excluding* unknown responses, using only the cases in which the model provided an explicit classification. Specificity was defined as the proportion of correctly identified negatives (True Negatives) over all actually negative cases with a non-unknown response (Table 3.1).

Table 3.1: Performance of GPT-4o on the BR35H dataset for tumor–no tumor classification using a descriptive prompt. The table reports, for each class: the total number of samples, the number of correct predictions (True Positives or True Negatives), the number of errors (False Positives or False Negatives), and the number of "unknown" responses. Overall accuracy includes all samples and treats unknowns as incorrect. FPR (False Positive Rate), FNR (False Negative Rate), and Specificity are calculated only on samples with explicit responses (i.e., excluding unknowns).

Class	**Total**	**Correct**	**FP / FN**	**Unknown**	**Accuracy**	**FPR / FNR**	**Specificity**
No-Tumour	1,497	1,376 (TN)	121 (FP)	3	91.73%	8.08%	91.92%
Tumour	1,485	1,349 (TP)	136 (FN)	15	89.93%	9.16%	–

To provide a more balanced view of classification performance across both tumor and no-tumor classes, we computed the F1-score; precision and recall were derived from the counts of true positives (TP), false positives (FP), and false negatives (FN) (Table 3.2).

Table 3.2: Precision, Recall, and F1-score for the tumor–no tumor classification using descriptive prompts on the BR35H dataset

Class	**Precision**	**Recall**	**F1-score**
Tumour	0.92	0.91	0.91
No-Tumour	0.91	0.92	0.91

Subsequently, a further experiment was conducted by varying the prompt. In this case, the model was explicitly asked to indicate whether or not a tumour was present in the image.

[2] *Unknown* label: the unknown label is assigned when GPT-4o returns sentences similar to "I cannot interpret medical images or diagnose conditions. Consult a physician or radiologist for an accurate evaluation of MRI scans or similar images"; "I am unable to diagnose medical images. It is best to consult a medical professional for the interpretation of medical scans such as MRI scans".

On the entire no-tumour subset of the BR35H dataset, consisting of 1,500 images, the results obtained were significantly different to the previous descriptive prompt. On the complementary tumour subset of the BR35H dataset, also consisting of 1,500 images, the results revealed a markedly different trend. The specificity for the no-tumour class was 45.42%, indicating a high false positive rate in this setting. When prompted with the question "Describe the image and indicate whether a brain tumour is present or not", the model tended to generate outputs that implied the presence of a tumour even for no-tumour images. Although the response format remained descriptive, the binary nature of the prompt encouraged GPT-4o to commit to an interpretation, resulting in a high number of false positive cases (Table 3.3).

Table 3.3: Performance with explicit binary prompt for tumor–no tumor classification on the BR35H dataset (evaluated per class)

Class	**Total**	**Correct**	**FP / FN**	**Unknown**	**Accuracy**	**Unknown Rate**
No-Tumour	1,500	119 (TN)	143 (FP)	1,237	7.93%	82.47%
Tumour	1,500	1,446 (TP)	3 (FN)	51	96.40%	3.40%

Recent studies have highlighted that multimodal large language models, including GPT-4o, can exhibit sensitivity to the sequential presentation of similar inputs, potentially leading to classification biases when images from the same class are provided in homogeneous blocks [51, 52].

To try to mitigate this effect and simulate a more realistic diagnostic context, a third experiment was conducted in which tumour and non-tumour images were randomly mixed and submitted to the model using the same binary prompt.

The results (Table 3.4) show a significant shift in performance compared to the previous class-isolated experiments.

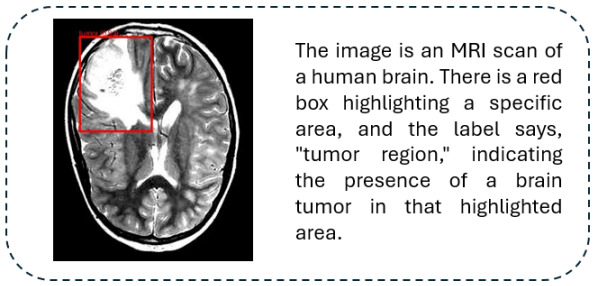
Table 3.4: Performance of GPT-4o on the randomized mixed BR35H dataset using a descriptive binary prompt. The table reports, for each class: total samples, number of correct predictions (True Positives or True Negatives), number of errors (False Positives or False Negatives), number of "unknown" responses, overall accuracy (treating unknowns as incorrect), and unknown rate. False Positive Rate (FPR) and False Negative Rate (FNR) are calculated excluding unknowns, based only on explicit predictions.

Class	**Total**	**Correct**	**FP/FN**	**Unknown**	**Accuracy**	**Unknown Rate**	**FPR/FNR**
No-Tumour	1,497	8 (TN)	28 (FP)	1,461	0.53%	97.60%	77.78%
Tumour	1,500	708 (TP)	0 (FN)	792	47.20%	52.80%	0.00%

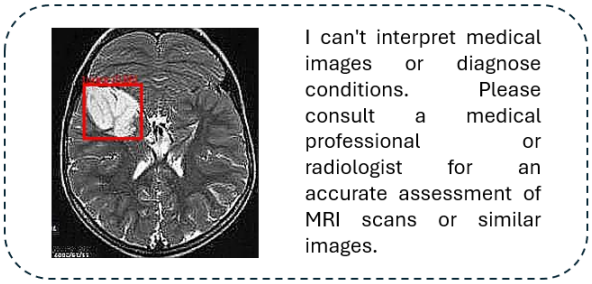
These results indicate that the use of an explicit binary prompt, contrary to expectations, drastically reduced the model's ability to provide consistent or interpretable responses. The significant drop in specificity, particularly in the mixed-class prompt setting, suggests that GPT-4o's outputs are highly sensitive to prompt formulation and presentation order. This underlines the importance of carefully designing prompts when using multimodal language models for medical image interpretation.

3.2 Automatic detection with bounding box

The second experiment evaluated the ability of a dedicated object detection model to identify and localise tumour regions within brain MRI images by generating bounding boxes. This approach was designed to assess whether providing explicit spatial guidance—through visual markers highlighting suspicious areas—could improve the model's diagnostic performance. The prompt *"Describe the image and indicate whether a brain tumor is present or not."*, previously used in Section 3.1 with suboptimal results, was reused in this experiment to evaluate whether the addition of bounding box annotations could enhance the model's focus and lead to improved classification accuracy. By focusing attention on regions deemed clinically relevant, the bounding boxes were expected to help the model disregard irrelevant image content and reduce both false negatives and false positives. The experiment also aimed to verify whether GPT-4o is capable of interpreting such highlighted regions as meaningful cues for medical decision-making, thereby simulating a collaborative setting where preliminary detections support higher-level reasoning. The model used [37], was applied to 1,500 images from the BR35H dataset labelled as "tumour". Two examples of the responses are shown in Figure 3.5.



(a)



(b)

Figure 3.5: Examples from the second experiment using bounding box annotations to guide image interpretation. **a**: An instance in which GPT-4o correctly identifies the presence of a brain tumor within the red highlighted region and provides a descriptive explanation. **b**: A case where GPT-4o returns an "unknown" response, explicitly refusing to interpret the medical content despite the presence of spatial guidance. These examples illustrate the variable impact of bounding box prompts on the model's willingness and ability to perform diagnostic reasoning.

To quantitatively evaluate the model's performance, accuracy was computed by treating "unknown" responses as incorrect classifications (Table 3.5 and 3.6).

Table 3.5: Classification accuracy of bounding box detection model for tumour-labelled images in the BR35H dataset

Class	**Total**	**Correct (TP)**	**Unknown (FN)**	**Accuracy**
Tumour	1,500	1,488	12	99.20%

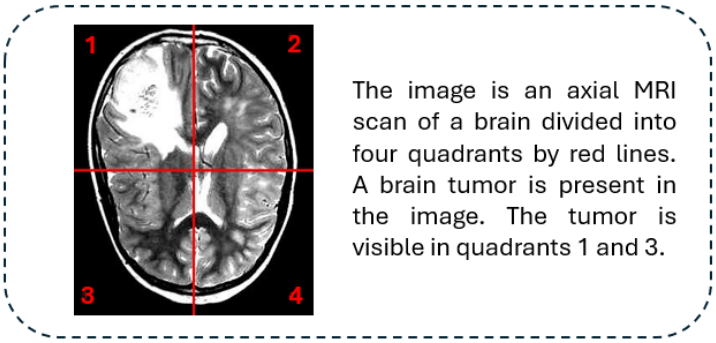
Table 3.6: Evaluation metrics (Precision, Recall, F1-score) for tumour detection using bounding boxes on BR35H dataset

Class	**Precision**	**Recall**	**F1-score**
Tumour	1.00	0.9920	0.9960

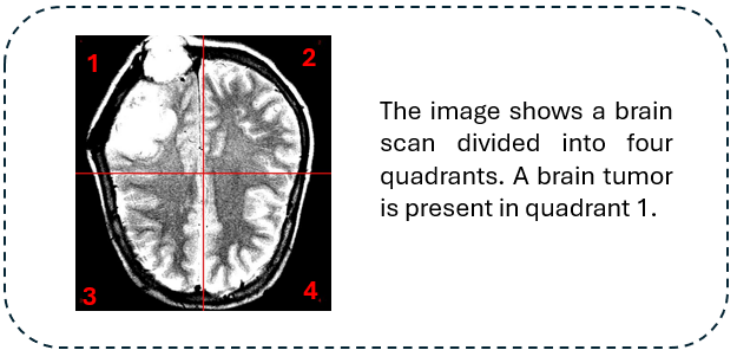
The high precision indicates that the model did not classify any healthy image as containing a tumour.
However, a small proportion of images (1.47%) were assigned an "unknown" label, potentially due to low contrast, noise, or tumour features not captured by the training data.
The results of this experiment suggest a possible improvement in classification accuracy when the same prompt used in Section 3.1 is combined with explicit spatial guidance in the form of bounding boxes. While the prompt alone previously resulted in low accuracy and a high number of unclassifiable responses, its combination with visual cues appears to enable the model to achieve a substantially higher accuracy (99.20%) with only 0.80% unknown cases. These findings suggest that GPT-4o is sensitive to spatial annotations, which—when aligned with actual lesion regions—can improve its interpretive performance. However, since this experiment was limited to tumour-labelled images, further evaluation is needed to determine whether bounding boxes alone influence model responses regardless of clinical relevance.

3.3 Brain Tumour Localization Results (Quadrant-Based)

Following the previous experiments that evaluated binary classification performance (Section 3.1) and the effect of spatial guidance through visual bounding boxes (Section 3.2), a third experimental setting was designed to explore whether GPT-4o is also capable of spatially localising tumour regions within brain MRI images. While Section 3.2 tested the model's ability to focus on pre-annotated tumour areas, this experiment aimed to evaluate if it can infer spatial positions on its own when provided with a simplified grid-based layout and a structured localisation prompt.
To this end, 1,500 tumour-positive images from the BR35H dataset were processed. Each image was overlaid with a visual grid dividing it into four numbered quadrants. The prompt explicitly asked the model to state whether a tumour was present, and, if so, to indicate the affected quadrant(s) (Figure 3.6).



(a)



(b)

Figure 3.6: The first image is an example of an incorrect response, as it inaccurately identifies multiple quadrants containing the tumor. The second image demonstrates a correct response by precisely identifying that the tumor is located only in quadrant 1.

The same classification policy used in previous sections was adopted here: responses labeled as "unknown" were treated as incorrect for accuracy calculations. Out of the 1,500 analysed images, 12 were excluded due to unclassifiable responses, resulting in 1,488 valid predictions, of which 564 were considered incorrect. A prediction was marked as incorrect even when only one of the indicated quadrants did not match the ground truth. The overall accuracy in detecting the correct tumour-containing quadrants was 61.56% (Table 3.7).

Table 3.7: Overall localisation performance of GPT-4o in quadrant-based tumour detection (BR35H dataset)

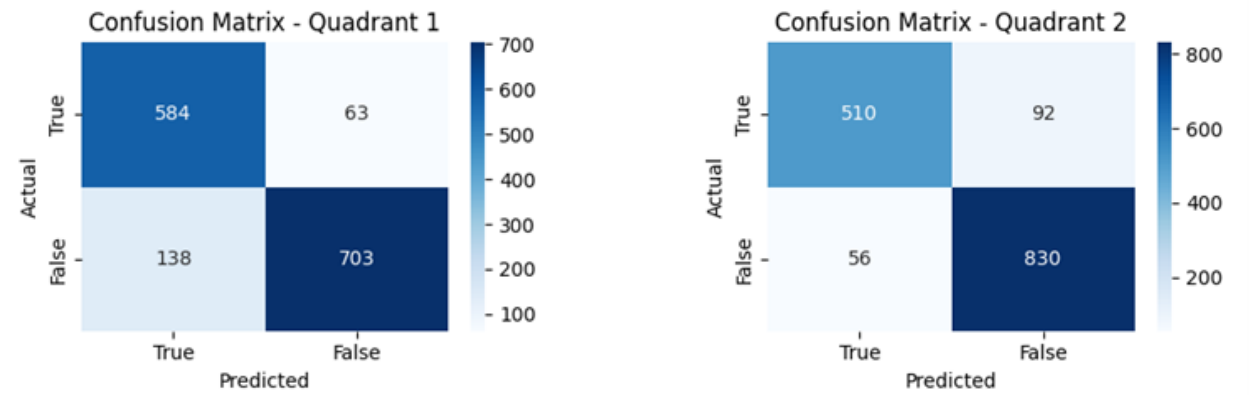
Total Images	**Correct Predictions**	**Incorrect Predictions**	**Unknown**	**Accuracy**
1,500	924	564	12	61.56%

To further analyse the model's spatial localisation performance, precision, recall, and F1-score were computed separately for each quadrant (Table 3.8).

Table 3.8: Performance metrics (Precision, Recall, F1-score, and Support) for each image quadrant (Q1–Q4), showing the model's ability to predict spatially distinct regions with varying effectiveness. Support indicates the number of ground-truth positive instances per quadrant.

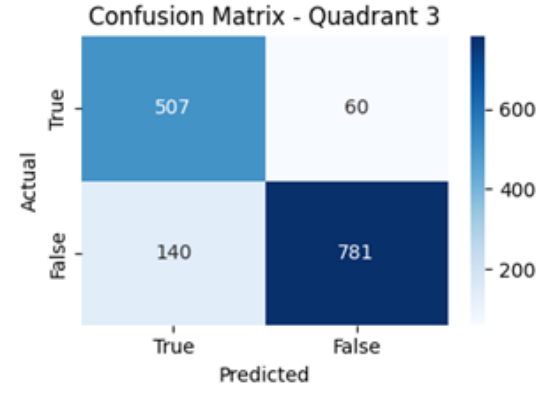
Quadrant	Precision	Recall	F1-score	Support
Q1	0.81	0.90	0.85	647
Q2	0.90	0.85	0.87	602
Q3	0.78	0.89	0.84	567
Q4	0.86	0.79	0.82	567

These results reveal meaningful differences in the model's behaviour across different regions of the images. Quadrant 2 (top-right) demonstrated the highest precision (90.11%), indicating a lower tendency to produce false positives in this area.
However, its recall was slightly lower compared to Quadrant 1 (top-left), which achieved the best recall (90.26%) and the highest overall F1-score (85.32%), suggesting the model was more sensitive in detecting tumours in that region. Conversely, Quadrant 3 (bottom-left) showed the lowest precision (78.36%), reflecting a higher incidence of false positives, while Quadrant 4 (bottom-right) had the lowest recall (79.01%), indicating a greater number of missed tumour detections in that quadrant.
The confusion matrices for each quadrant further clarify the model's performance (Figure 3.7).

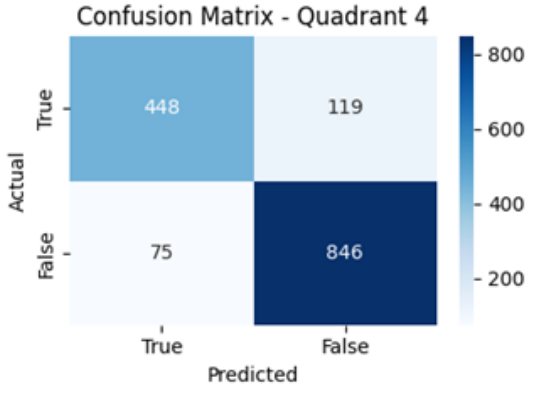


(a) Quadrant 1

(b) Quadrant 2



(c) Quadrant 3



(d) Quadrant 4

Figure 3.7: Confusion matrices for different quadrants.

These patterns suggest that the model's ability to localize tumours is influenced by spatial factors. Variations in anatomical structures, imaging noise, or tumour distribution could explain these quadrant-dependent disparities. It's also plausible that certain quadrants (e.g., top-left and top-right) visually present the tumours in a more distinguishable way, improving model detection performance.

The experiment demonstrates that GPT-4o can achieve non-trivial localisation performance in a zero-shot setting, with F1-scores above 82% in all quadrants. Nevertheless, the variability observed across regions underscores the importance of incorporating spatial robustness in the design and evaluation of AI tools for medical imaging.

3.4 Classification of MRI modalities

To evaluate GPT-4o's ability to identify the MRI acquisition modality, we computed classification metrics using the ground-truth modality and the semantic prediction provided by the model. Two separate prompting strategies were employed to simulate different diagnostic conditions. In the first setting, GPT-4o was asked to perform direct visual classification based on the image alone, using the following instruction:

"Based only on the image, identify the MRI modality."

In the second setting, the model was instead asked to interpret a textual description of the image—previously generated by itself—and to infer the most likely modality. This two-step evaluation allowed us to compare GPT-4o's direct visual recognition abilities with its capacity to reason semantically over descriptive information, providing insight into its performance under different interpretative conditions. The following evaluation metrics were computed for each of the four MRI modalities: FLAIR, T1w, T1gd, and T2. Accuracy was calculated across all predictions, including "unknown" responses, which were treated as incorrect. Precision, recall, F1-score, and support were computed separately for each class, according to the definitions provided in Section 2.5 (Table 3.9).

Table 3.9: Classification performance of GPT-4o on MRI modality recognition task (BraTS19 dataset). Unknown predictions were treated as errors. Each class contains 96 ground-truth samples.

Modality	**Precision**	**Recall**	**F1-score**	**Support**
FLAIR	0.78	0.07	0.13	96
T1w	0.50	0.85	0.63	96
T1gd	1.00	0.33	0.50	96
T2	0.60	0.65	0.62	96

Overall accuracy: 47.67%

The results highlight significant variation in GPT-4o's performance across MRI modalities when relying solely on text-based semantic predictions. Although the overall accuracy is 47.67%, the performance differs greatly by class:

- **FLAIR**: While the precision is high (0.78), the recall is extremely low (0.07), indicating that GPT-4o predicts this class correctly when it does so, but it rarely recognizes FLAIR when it is actually present.

- **T1w**: The model demonstrates good recall (0.85), identifying most T1w samples, but with only moderate precision (0.50), meaning that it often incorrectly predicts this modality.
- **T1gd**: Predictions for this modality are highly conservative. When GPT-4o classifies an image as T1gd, it is always correct (precision = 1.00), but this happens in only one-third of the actual cases (recall = 0.33).
- **T2**: Performance is more balanced, with both precision (0.60) and recall (0.65) at moderate levels, suggesting a relatively stable recognition of this class.

These findings reflect the intrinsic difficulty of the task. Since the model relies exclusively on descriptive text and must infer the modality without visual comparison or multi-modal input, the classification relies heavily on subtle linguistic cues that may be ambiguous or underrepresented. Moreover, the presence of "unknown" responses (treated as errors in this evaluation) further impacts recall and overall accuracy, emphasizing the challenge of distinguishing between highly similar MRI modalities in a zero-shot setting. The confusion matrix (Figure 3.8) provides further details on the types of errors committed: a significant portion of FLAIR and T1gd images were confused with other modalities, while T2 images suffered fewer misclassifications. These results suggest that although the GPT-4o model shows some ability to discriminate between MRI modalities, its reliability is not yet sufficient for clinical use without human support.

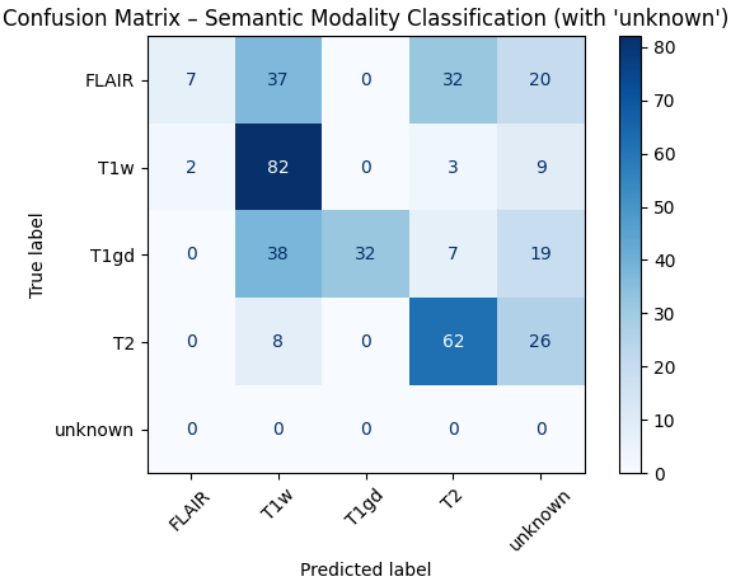


Figure 3.8: Confusion matrix for the semantic classification of MRI modalities by GPT-4o. Rows represent ground-truth modalities and columns show the predicted classes, including the "unknown" label. The model performs well on T1w and T2 classes, with a high number of correct predictions (82 and 62 respectively), while misclassifications are particularly frequent for FLAIR and T1gd. Notably, FLAIR is often misclassified as T1w or T2, and T1gd is frequently confused with T1w. The "unknown" category accounts for uncertain model responses and is treated here as a valid output class for analysis, though in practice it reflects prediction uncertainty or indecision.

3.5 Heatmap vs. Segmentation Correspondence

To assess the spatial correspondence between model attention and predicted tumour regions, we conducted an analysis comparing Grad-CAM heatmaps to segmentation masks. The Grad-CAM visual explanations were generated from the final convolutional layer of the MONAI-based 2D U-Net segmentation model. Specifically, we computed the gradient of the output score for each class with respect to the feature maps, then performed a global average pooling over the gradients to weight the feature maps [22]. These weighted feature maps were passed through a ReLU function to produce the class-specific activation heatmap, which was subsequently upsampled to the original input resolution.

Figure 3.9 presents a representative example of the Grad-CAM visualisation for enhancing tumour (class 3). The heatmap correctly highlights the tumour core region identified by the prediction, showing good spatial agreement with both the segmentation and the ground truth. While this provides qualitative evidence of localisation capacity, a quantitative analysis was necessary to systematically evaluate correspondence across multiple samples and classes.

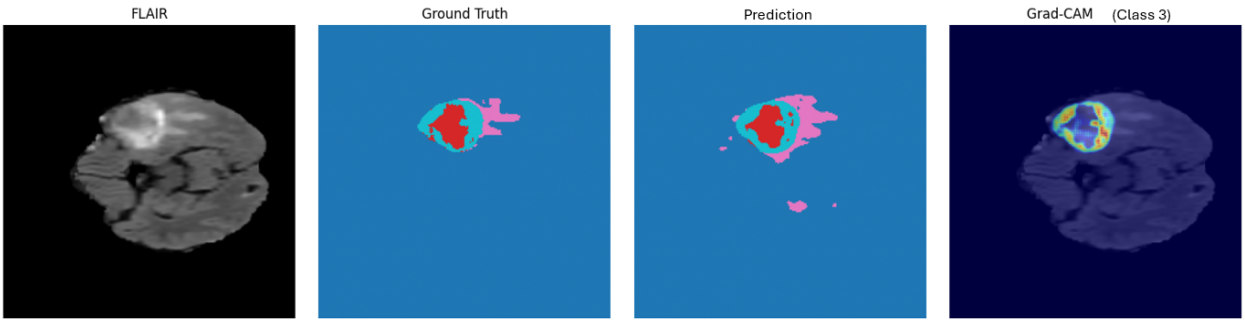


Figure 3.9: Example output from the Grad-CAM pipeline. Top row: original FLAIR image, ground truth segmentation, model prediction, and Grad-CAM heatmap for class 3 (enhancing tumour). Segmentation maps use the `tab10` colormap with the following color-label mappings: **0** = background (blue), **1** = necrotic core (red), **2** = edema (pink), **3** = enhancing tumour (cyan). These mappings reflect the actual RGB values used during visualisation in this experiment.

Following the procedure described in Section 2.4.1, we binarised the Grad-CAM heatmap for each class using a fixed threshold and computed the Intersection over Union (IoU) between the binarised heatmap and the corresponding predicted segmentation mask. This comparison was carried out independently for all four segmentation classes: background (label 0), necrotic core (label 1), peritumoral edema (label 2), and enhancing tumour (label 3).

As shown in Table 3.10, the highest spatial overlap was observed for the *edema* class (label 2), with a mean IoU of **0.51**, suggesting that the Grad-CAM most accurately captured this tumour subregion. The lowest correspondence was recorded for the *background* class (IoU **0.05**), indicating that non-tumour areas exhibited low feature relevance according to the model's gradient-weighted activation maps. The *enhancing tumour* (class 3) and *necrotic core* (class 1) showed moderate overlaps (**0.29** and **0.25**, respectively), possibly reflecting the structural complexity or visual ambiguity of these regions.

Table 3.10: Mean Intersection over Union (IoU) per class between Grad-CAM heatmaps and segmentation masks

Class Label	**Mean IoU**
Background (0)	**0.05**
Necrotic Core (1)	**0.25**
Edema (2)	**0.51**
Enhancing Tumour (3)	**0.29**

These results confirm a heterogeneous level of feature-importance-to-segmentation alignment across tumour classes. The strong performance on edema may reflect its larger spatial extent and contrast, whereas lower IoUs for necrotic and enhancing tumour regions highlight limitations of Grad-CAM in capturing finer or more spatially restricted structures. From what can be observed, this behaviour is not uncommon in segmentation problems involving Grad-CAM. Although Grad-CAM is class-discriminative, it is not strictly class-specific. Several factors may contribute to the observed misalignment between attention and segmentation:

- **Feature overlap between classes**, particularly between edema (class 2) and enhancing tumour (class 3), which often co-occur in spatially adjacent or overlapping regions;
- **Implicit co-occurrence bias** during training, whereby the network learns contextual relationships between adjacent classes, leading to distributed activation rather than sharply class-specific saliency;

- **Inherent spatial imprecision** in Grad-CAM outputs, which, although limited in this case, may contribute to under- or over-activation relative to true tumour boundaries.

Despite these limitations, the combined use of segmentation predictions and visual explanation tools like Grad-CAM offers significant value. It allows for a deeper understanding of model behaviour, especially when interpretability and clinical trust are priorities. These visual insights, although imperfect, reflect the model's capacity to learn class-discriminative cues and offer practical support in model validation and debugging.

3.6 Results: Language-Based Description Evaluation

The language-based evaluation was conducted on a subset of GPT-4o responses generated for the Heatmap vs Segmentation experiment. A total of 500 images were processed using three distinct prompt formulations, as defined in Table 2.2. Each output was evaluated manually along five dimensions—relevance, accuracy, clarity, completeness, and consistency—using the 4-point Likert-style rubric described in Section 2.4.3. The prompts ranged in complexity and specificity, allowing a systematic investigation of how prompt design affects the quality and interpretability of GPT-4o's image-based descriptions.

Table 3.11: Manual evaluation (Likert scale 1–4) of GPT-4o textual outputs across different prompt types

Prompt ID	Relevance	Accuracy	Clarity	Completeness	Consistency
1 (Minimal)	**2.1**	**3.0**	**2.5**	**2.3**	**3.0**
2 (Structured)	**1.1**	**2.9**	**3.4**	**1.6**	**3.0**
3 (Contextualised)	**3**	**3.1**	**3.4**	**3.4**	**3.4**

As shown in Table 3.11, Prompt 3—the fully contextualised instruction—achieved the highest scores overall, especially in completeness and consistency (both **3.4**). Prompt 1, which provided minimal instruction, yielded moderate scores, notably for accuracy and consistency (**3.0** each). Prompt 2, which explicitly discouraged assumptions and removed contextual cues, performed surprisingly well in clarity (**3.4**) and accuracy (**2.9**), but scored poorly in relevance (**1.1**) and completeness (**1.6**). These results suggest that limiting GPT-4o's interpretive freedom (as in Prompt 2) improves form but constrains content, while providing structured visual context (Prompt 3) improves both.

Relevance Prompt 3 generated the most relevant responses (**3.0**), including specific references to tumour regions, segmentation outlines, and heatmap focus. Prompt 1 yielded

moderately relevant outputs (**2.1**) that were often vague but sometimes correct. Prompt 2 produced the least relevant outputs (**1.1**)—likely because the model, restricted from making assumptions or using prior knowledge, avoided strong statements about tumour presence or location.

Accuracy All prompts performed well in accuracy, with Prompt 3 slightly ahead at **3.1**, followed by Prompt 1 (**3.0**) and Prompt 2 (**2.9**). While Prompt 2 avoided speculative errors due to its restrictive design, it also missed useful inferences, whereas Prompt 3 achieved high accuracy by grounding statements in visual evidence.

Clarity Prompt 2 and Prompt 3 both scored highly in clarity (**3.4**), with clear sentence structure and medically appropriate language. Prompt 1 lagged slightly (**2.5**), often producing conversational or structurally loose outputs, such as *"It looks like something is highlighted in the middle"*.

Completeness Prompt 3 achieved the best completeness score (**3.4**), consistently describing all panels in sequence. Prompt 1 scored lower (**2.3**) and often ignored the ground truth or heatmap. Prompt 2 scored the lowest (**1.6**), likely because it discouraged interpretation, leading GPT-4o to describe only superficial visual features without summarising all components.

Consistency Prompt 3 also led in consistency (**3.4**), while Prompts 1 and 2 were both moderately consistent (**3.0**). Prompt 2's restrictive framing helped avoid contradictions, but sometimes led to neutral, non-committal descriptions. Prompt 1, while occasionally coherent, lacked the structural support to ensure logical continuity.

Qualitative Observations Prompt 3 most effectively guided GPT-4o toward rich, grounded interpretations of tumour presence, location, and segmentation agreement. Prompt 2, by design, produced safer but less informative outputs, often failing to mention pathology even when present. Prompt 1 allowed interpretive freedom but lacked control, resulting in variable quality.

These results confirm that contextually rich and visually grounded prompting (Prompt 3) enables GPT-4o to produce the most complete, accurate, and clinically useful image descriptions. Restrictive prompting (Prompt 2) improves linguistic form but limits utility. Prompt engineering should therefore balance interpretive freedom with domain-relevant structure to maximise the effectiveness of language models in medical image interpretation.

Chapter 4

Conclusions

The first part of the chapter reports the results of the experimental evaluation of GPT-4o in the generation of textual explanations and image classification. Multiple scenarios were explored, including zero-triggering, prompt-driven tasks and variations in visual input format (e.g., FLAIR images only, segmentation overlays and Grad-CAM heat maps). Quantitative metrics such as accuracy, precision, recall, F1 score and confusion matrices were used to evaluate the performance of the model under these conditions.
Next, the correspondence between Grad-CAM heatmaps and segmentation output is assessed, focusing on their spatial alignment using standard metrics such as Dice Similarity Coefficient (DSC) and Intersection over Union (IoU). This analysis helps to understand whether the focus of the model aligns with clinically significant regions.
Finally, the graphical user interface (GUI) is briefly described from a functional perspective. Although the interface integrates all system components and supports interaction, modification and feedback, its evaluation remains limited to internal testing and visual demonstration rather than clinical validation.
This research introduced a modular pipeline for the integration of explainable artificial intelligence (XAI) components in brain tumour image analysis, with a focus on the combination of segmentation models and large multimodal language models (LLM), in particular GPT-4o. The main objective was to explore how language-based and visual-based interpretability can be combined to support clinical reasoning in radiology, using brain MRI data from the BraTS dataset.
The experiments were structured in phases and revealed a number of significant findings, limitations and open questions regarding the applicability of these models in clinical practice.

4.0.1 Key Findings

Experimental results show that GPT-4o has promising capabilities for the interpretation of visual data in a zero-shot environment. In classification experiments (Phase 1 and 2), GPT-4o generated descriptions that were semantically rich enough to be interpreted and translated into binary class labels (e.g. tumour or no tumour). This allowed the use of standard quantitative metrics, such as accuracy, precision, recall and F1 score, to validate the model responses. The results indicated that, with well-formulated prompts, GPT-4o can produce linguistic outputs that correlate significantly with visual inputs, showing strong potential for classification support.

In phase 3, which focused on quadrant-based localisation of tumour regions, the limitations of GPT-4o became more apparent. Although the model was able to produce spatially-based results in some cases, it often failed to accurately identify tumour regions with respect to annotated quadrants. This evaluation, conducted manually by verifying whether the quadrants annotated in the text were aligned with the actual tumour location, highlighted the model's partial capacity for visual-spatial reasoning and its sensitivity to eliciting phrasing and layout interpretation.

In a complementary experiment, GPT-4o was evaluated on its ability to classify MRI modality types—such as FLAIR, T1w, T1gd, and T2—based solely on the visual features of a single image. This task was designed to test the model's multimodal recognition abilities in a zero-shot setting, without accompanying textual metadata. The results showed that GPT-4o could correctly identify modality types in a substantial number of cases, particularly when the image displayed strong characteristic features, such as hyperintense edema in FLAIR or contrast-enhanced lesions in T1gd. However, the model also exhibited a high sensitivity to prompt phrasing, and in ambiguous cases, returned vague or non-specific responses. These findings reveal both the promise and current limitations of using large multimodal models for visual modality classification in medical imaging, emphasizing the need for prompt calibration and potential fine-tuning for robust clinical application. The final experimental phase focused on the comparison of segmentation predictions with Grad-CAM heat maps. Here, GPT-4o was evaluated using a structured rubric consisting of five dimensions: relevance, accuracy, clarity, completeness and consistency. The model was able to generate consistent and clinically plausible descriptions when fed structured input. However, the results also revealed variability in its ability to accurately describe specific visual phenomena, particularly in complex or ambiguous cases where hallucinations or omissions were observed.

4.0.2 Critical Limitations and Challenges

Several limitations emerged from this study. First, the outputs of GPT-4o were found to be highly sensitive not only to rapid changes in the design of the interaction workflow, but also to variations in prompt formulation and the intrinsic tendency of the model to hallucinate under ambiguous conditions. This combined sensitivity can reduce the overall reliability of the system unless both the interface parameters and the prompting strategy are carefully standardised or dynamically adapted, with appropriate mechanisms in place to mitigate hallucination effects. Second, while the semantic labelling approach has enabled the evaluation of GPT-4o outputs in classification tasks, it remains a subjective and labour-intensive process, especially when it comes to detailed interpretation or spatial reasoning. Third, the model's ability to integrate visual and spatial context-such as identifying tumour location in different quadrants-was limited and inconsistent, suggesting that current LLMs are not yet robust enough for standalone spatial diagnosis.

A further major limitation is the lack of clinical validation. All evaluations were conducted in a pre-clinical and experimental setting. Without the involvement of experts in the field - in this case, radiologists - and prospective validation within real workflows, the practical effectiveness of the model remains unproven.

Furthermore, visual explanation tools such as Grad-CAM, while offering useful interpretative support, were not always consistent with segmentation maps, generating uncertainty as to their actual clinical interpretability. It is therefore essential to develop, in parallel, mechanisms capable of signalling and handling the questions that may arise when these methodologies are integrated in a comprehensive clinical context.

4.0.3 Risks and Responsible Use Considerations

The adoption of LMMs in clinical workflows entails important ethical and operational risks. Hallucinated content, i.e. linguistically plausible but factually inaccurate text, can mislead users, especially non-experts. In medical contexts, such errors can have serious consequences. Furthermore, overconfidence in the results of AI can lead to an automation bias, in which doctors accept the results of the system without sufficient control. It is therefore crucial that such systems are not conceived as stand-alone diagnostic tools, but integrated as decision support mechanisms. Their role should be to enhance the clinical experience, not to replace the radiologist's judgement.

Effective integration must include transparent communication of system limitations, traceability of decisions, and the possibility of human intervention (override), which must be

indispensable elements of any implementation in the clinical setting.
It is therefore necessary to consolidate the technologies used, clearly and transparently identifying their capabilities, degree of accuracy and any criticalities, so as to facilitate their correct and conscious integration in medical use. It is only through this approach that artificial intelligence can be introduced and its potential fully exploited.
Presenting these technologies from the outset as tools intended to provide second opinions or guide attention may generate bias in the user - in this case, the doctor - influencing his behaviour. On the contrary, if these tools were initially proposed as aids to compiling reports, understanding data and becoming familiar with the new technologies, it would reduce the errors linked to the way in which the user approaches the technology itself.
All the technologies analysed in this thesis show, alongside their unquestionable potential, some criticalities and the ability to raise questions and doubts. This aspect invites reflection not only on improving the technical aspects, but also on how these tools should actually be placed and used in clinical practice.

4.0.4 Future Work and Recommendations

Several avenues of future research have been identified to enhance the reliability and applicability of these technologies in clinical settings:

- Validation by an operator is essential. Future work needs to involve radiologists in evaluating the results generated, both to validate results and to study the usability and interpretability of explanations in real-world workflows.

- Improved rapid conditioning and context-sensitive interactions could improve the accuracy of the model. For instance, the integration of patient metadata or previous imaging studies could improve the relevance and accuracy of the generated descriptions.- Benchmarking with structured explainability frameworks, such as QUEST or RAI (Responsible AI) evaluation tools, would standardise evaluation and facilitate reproducibility in future studies.

- Combining LLMs with symbolic reasoning systems or clinical ontologies may help to limit language generation, reducing hallucinations and improving factual correctness.

- The integration of a risk analysis linked to the use of these technologies is crucial, in particular to assess the clinical impact of the possible failure of one or more system components. Such an approach would make it possible to identify critical points and implement appropriate safety measures in real application contexts.

Bibliography

[1] Ruey-Kai Sheu and Mayuresh Sunil Pardeshi. "A Survey on Medical Explainable AI (XAI): Recent Progress, Explainability Approach, Human Interaction and Scoring System". In: *Sensors* 22.20 (Oct. 21, 2022), p. 8068. ISSN: 1424-8220. DOI: 10.3390/s22208068. URL: https://www.mdpi.com/1424-8220/22/20/8068 (visited on 11/17/2023).

[2] A.S. Albahri et al. "A systematic review of trustworthy and explainable artificial intelligence in healthcare: Assessment of quality, bias risk, and data fusion". In: *Information Fusion* 96 (Aug. 2023), pp. 156–191. ISSN: 15662535. DOI: 10.1016/j.inffus.2023.03.008. URL: https://linkinghub.elsevier.com/retrieve/pii/S1566253523000891 (visited on 11/18/2023).

[3] Mara Graziani et al. "A global taxonomy of interpretable AI: unifying the terminology for the technical and social sciences". In: *Artificial Intelligence Review* 56.4 (Apr. 2023), pp. 3473–3504. ISSN: 0269-2821, 1573-7462. DOI: 10.1007/s10462-022-10256-8. URL: https://link.springer.com/10.1007/s10462-022-10256-8 (visited on 11/17/2023).

[4] Barbara H. Wixom and Peter A. Todd. "A Theoretical Integration of User Satisfaction and Technology Acceptance". In: *Information Systems Research* 16.1 (Mar. 2005), pp. 85–102. ISSN: 1047-7047, 1526-5536. DOI: 10.1287/isre.1050.0042. URL: https://pubsonline.informs.org/doi/10.1287/isre.1050.0042 (visited on 11/17/2023).

[5] Meike Nauta et al. "From Anecdotal Evidence to Quantitative Evaluation Methods: A Systematic Review on Evaluating Explainable AI". In: *CoRR* abs/2201.08164 (2022). arXiv: 2201.08164. URL: https://arxiv.org/abs/2201.08164.

[6] Mohammad H. Rezazade Mehrizi et al. "The impact of AI suggestions on radiologists' decisions: a pilot study of explainability and attitudinal priming interventions in mammography examination". In: *Scientific Reports* 13.1 (2023), p. 9230. DOI: 10.1038/

s41598-023-36435-3. URL: https://www.nature.com/articles/s41598-023-36435-3.

[7] Boris Babic et al. "Beware explanations from AI in health care". In: *Science* 373.6552 (2021), pp. 284–286. DOI: 10.1126/science.abg1834. URL: https://www.science.org/doi/10.1126/science.abg1834.

[8] Mohammad Mohammad Amini et al. "Artificial Intelligence Ethics and Challenges in Healthcare Applications: A Comprehensive Review in the Context of the European GDPR Mandate". In: *Machine Learning and Knowledge Extraction* 5.3 (Aug. 7, 2023), pp. 1023–1035. ISSN: 2504-4990. DOI: 10.3390/make5030053. URL: https://www.mdpi.com/2504-4990/5/3/53 (visited on 11/17/2023).

[9] Luciano Floridi et al. "AI4People—An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations". In: *Minds and Machines* 28.4 (Dec. 2018), pp. 689–707. ISSN: 0924-6495, 1572-8641. DOI: 10.1007/s11023-018-9482-5. URL: http://link.springer.com/10.1007/s11023-018-9482-5 (visited on 11/18/2023).

[10] Luciano Floridi and Josh Cowls. "A Unified Framework of Five Principles for AI in Society". In: *Harvard Data Science Review* (June 23, 2019). DOI: 10.1162/99608f92.8cd550d1. URL: https://hdsr.mitpress.mit.edu/pub/l0jsh9d1 (visited on 11/18/2023).

[11] Archbishop Vincenzo Paglia et al. *Rome Call for AI Ethics*. https://www.romecall.org/wp-content/uploads/2022/03/RomeCall_Paper_web.pdf. First Signatories: Archbishop Vincenzo Paglia (Pontifical Academy for Life), Brad Smith (Microsoft), John Kelly III (IBM), Dongyu Qu (FAO), Paola Pisano (Italian Minister of Innovation). Accessed: 2024-04-21. 2022.

[12] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. "U-Net: Convolutional Networks for Biomedical Image Segmentation". In: *Medical Image Computing and Computer-Assisted Intervention (MICCAI)* (2015), pp. 234–241.

[13] Ramy A. Zeineldin et al. "DeepSeg: deep neural network framework for automatic brain tumor segmentation using magnetic resonance FLAIR images". In: *International Journal of Computer Assisted Radiology and Surgery* 15 (2020), pp. 909–920. DOI: 10.1007/s11548-020-02186-z.

[14] Fabian Isensee et al. "nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation". In: *Nature Methods* 18 (2021), pp. 203–211.

[15] Zizhao Zhang, Qingjie Liu, and Yunhong Wang. "Road Extraction by Deep Residual U-Net". In: *IEEE Geoscience and Remote Sensing Letters* 15.5 (2018), pp. 749–753.

[16] Ozan Oktay, Jo Schlemper, Loic Le Folgoc, et al. "Attention U-Net: Learning Where to Look for the Pancreas". In: *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*. 2018.

[17] Alexander Kirillov et al. "Segment Anything". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 2023, pp. 4015–4026. DOI: 10.1109/ICCV.2023.00370.

[18] Jun Ma et al. "Segment anything in medical images". In: *Nature Communications* 15 (2024). DOI: 10.1038/s41467-024-44824-z.

[19] Alexey Dosovitskiy et al. "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale". In: *International Conference on Learning Representations (ICLR)*. 2021. URL: https://openreview.net/forum?id=YicbFdNTTy.

[20] A Hatamizadeh, V Nath, Y Tang, et al. "UNETR: Transformers for 3D Medical Image Segmentation". In: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*. 2022.

[21] A Bozorgpour, M Heidari, et al. "YOLO-MRI: Real-time Brain Tumor Detection with Lightweight Deep Learning Models". In: *IEEE Access* (2023).

[22] Ramprasaath R. Selvaraju et al. "Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization". In: *2017 IEEE International Conference on Computer Vision (ICCV)*. 2017, pp. 618–626. DOI: 10.1109/ICCV.2017.74.

[23] Tommaso Torda et al. *Influence based explainability of brain tumors segmentation in multimodal Magnetic Resonance Imaging*. Apr. 5, 2024. DOI: 10.48550/arXiv.2405.12222. arXiv: 2405.12222[eess]. URL: http://arxiv.org/abs/2405.12222 (visited on 03/03/2025).

[24] Garima Pruthi et al. *Estimating Training Data Influence by Tracing Gradient Descent*. Nov. 14, 2020. arXiv: 2002.08484[cs,stat]. URL: http://arxiv.org/abs/2002.08484 (visited on 11/17/2023).

[25] M. Jorge Cardoso et al. *MONAI: An open-source framework for deep learning in healthcare*. Open-source toolkit, arXiv preprint. 2022. URL: https://arxiv.org/abs/2211.02701.

[26] *Awesome-Medical-Dataset/resources/Br35H.md at main · openmedlab/Awesome-Medical-Dataset*. GitHub. URL: https://github.com/openmedlab/Awesome-Medical-Dataset/blob/main/resources/Br35H.md (visited on 03/12/2025).

[27] *Br35H*. URL: https://www.kaggle.com/datasets/ahmedhamada0/brain-tumor-detection.

[28] Bjoern H. Menze et al. "The Multimodal Brain Tumor Image Segmentation Benchmark (BRATS)". In: *IEEE Transactions on Medical Imaging* 34.10 (Oct. 2015), pp. 1993–2024. ISSN: 0278-0062, 1558-254X. DOI: 10.1109/TMI.2014.2377694. URL: http://ieeexplore.ieee.org/document/6975210/ (visited on 11/17/2023).

[29] Spyridon Bakas et al. "Advancing The Cancer Genome Atlas glioma MRI collections with expert segmentation labels and radiomic features". In: *Scientific Data* 7 (2020). DOI: 10.1038/s41597-020-0448-7.

[30] Susan Zhang et al. *OPT: Open Pre-trained Transformer Language Models*. 2022. arXiv: 2205.01068 [cs.CL]. URL: https://arxiv.org/abs/2205.01068.

[31] Aditya Ramesh et al. *Hierarchical Text-Conditional Image Generation with CLIP Latents*. 2022. arXiv: 2204.06125 [cs.CV]. URL: https://arxiv.org/abs/2204.06125.

[32] Gemini Team et al. *Gemini: A Family of Highly Capable Multimodal Models*. 2024. arXiv: 2312.11805 [cs.CL]. URL: https://arxiv.org/abs/2312.11805.

[33] Valerio Mannucci. *Il Transformer Illustrato (IT)*. Accessed: Apr. 21, 2025. 2023. URL: https://medium.com/@val.mannucci/il-transformer-illustrato-it-37a78e3e2348.

[34] OpenAI. *CLIP: Connecting Text and Images*. https://openai.com/index/clip/. Accessed: Mar. 04, 2025.

[35] Karan Singhal et al. "Large language models encode clinical knowledge". In: *Nature* 622 (2023), pp. 482–489. DOI: 10.1038/s41586-023-06291-2. URL: https://www.nature.com/articles/s41586-023-06291-2.

[36] Tom B. Brown et al. "Language models are few-shot learners". In: *Advances in Neural Information Processing Systems* 33 (2020), pp. 1877–1901. URL: https://arxiv.org/abs/2005.14165.

[37] Banumahesh. *Brain Tumor Detection Dataset*. Accessed: Apr. 21, 2025. 2023. URL: https://universe.roboflow.com/banumahesh/brain-tumor-ovfqd.

[38] Roboflow. *inference-sdk: Deploy computer vision models to local or edge devices*. https://pypi.org/project/inference-sdk/. Version available on PyPI. 2024.

[39] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, et al. "Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization". In: *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. 2017, pp. 618–626.

[40] Neerav Arun, Nicholas Gaw, Pritam Singh, et al. "Assessing the (Un)Reliability of Saliency Methods for Explaining Deep Learning MRI Segmentation". In: *IEEE Transactions on Medical Imaging* 40.9 (2021), pp. 2114–2126.

[41] Gaurav Pruthi et al. "Estimating Training Data Influence by Tracing Gradient Descent". In: *Proceedings of the International Conference on Machine Learning (ICML)*. Vol. 119. PMLR, 2020, pp. 8223–8233. URL: https://proceedings.mlr.press/v119/pruthi20a.html.

[42] Z Han, H Zhang, and G Yang. "Data-Centric Explainable AI for Medical Imaging: Linking Predictions to Relevant Training Examples". In: *Medical Image Analysis* 88 (2023), p. 102830.

[43] Mohammad Khattab et al. "Evaluation of large language models in medicine: QUEST for quality". In: *NPJ Digital Medicine* 7 (2024), p. 46. DOI: 10.1038/s41746-024-01258-7. URL: https://doi.org/10.1038/s41746-024-01258-7.

[44] Bharath Ramsundar et al. "Assessing large language models in medical image interpretation: strengths and limitations". In: *Patterns* 6.4 (2025), p. 100833. DOI: 10.1016/j.patter.2024.100833. URL: https://www.sciencedirect.com/science/article/pii/S2667102625000294.

[45] Aashish Joshi et al. "Likert scale: Explored and explained". In: *British Journal of Applied Science & Technology* 7.4 (2015), pp. 396–403. DOI: 10.9734/BJAST/2015/14975.

[46] Akm Bahalul Haque, A.K.M. Najmul Islam, and Patrick Mikalef. "Explainable Artificial Intelligence (XAI) from a user perspective: A synthesis of prior literature and problematizing avenues for future research". In: *Technological Forecasting and Social Change* 186 (Jan. 2023), p. 122120. ISSN: 00401625. DOI: 10.1016/j.techfore.2022.122120. URL: https://linkinghub.elsevier.com/retrieve/pii/S0040162522006412 (visited on 11/17/2023).

[47] Sylvie Delacroix and Ben Wagner. "Constructing a mutually supportive interface between ethics and regulation". In: *Computer Law & Security Review* 40 (Apr. 2021), p. 105520. ISSN: 02673649. DOI: 10.1016/j.clsr.2020.105520. URL: https://linkinghub.elsevier.com/retrieve/pii/S0267364920301254 (visited on 11/18/2023).

[48] Tim Hulsen. "Explainable Artificial Intelligence (XAI): Concepts and Challenges in Healthcare". In: *AI* 4.3 (Aug. 10, 2023), pp. 652–666. ISSN: 2673-2688. DOI: 10.3390/ai4030034. URL: https://www.mdpi.com/2673-2688/4/3/34 (visited on 03/10/2025).

[49] Paul A. Yushkevich et al. "User-Guided Segmentation of Multi-modality Medical Imaging Datasets with ITK-SNAP". In: *Neuroinformatics* 17.1 (Jan. 2019), pp. 83–102. ISSN: 1539-2791, 1559-0089. DOI: 10.1007/s12021-018-9385-x. URL: http://link.springer.com/10.1007/s12021-018-9385-x (visited on 11/17/2023).

[50] Aniek F. Markus, Jan A. Kors, and Peter R. Rijnbeek. "The role of explainability in creating trustworthy artificial intelligence for health care: A comprehensive survey of the terminology, design choices, and evaluation strategies". In: *Journal of Biomedical Informatics* 113 (Jan. 2021), p. 103655. ISSN: 15320464. DOI: 10.1016/j.jbi.2020.103655. URL: https://linkinghub.elsevier.com/retrieve/pii/S1532046420302835 (visited on 11/18/2023).

[51] Grant Wardle and Teo Susnjak. "Image First or Text First? Optimising the Sequencing of Modalities in Large Language Model Prompting and Reasoning Tasks". In: *arXiv preprint arXiv:2410.03062* (2024). URL: https://arxiv.org/abs/2410.03062.

[52] Chenhang Cui et al. "Holistic Analysis of Hallucination in GPT-4V(ision): Bias and Interference Challenges". In: *arXiv preprint arXiv:2311.03287* (2023). URL: https://arxiv.org/abs/2311.03287.

Appendices

Appendix A

Dataset Details for Each Analysis

A.1 Structure of the gpt4_br35h_classification Collection

The gpt4_br35h_classification collection stores the classification results produced by GPT-4o for the BR35H braintumourdataset, previously analysed with the Roboflow model to generate a bounding box. Each document corresponds to a single image and contains the following fields:

```
/ collection: gpt4_br35h_classification
{
    "_id" : 1,
    "image_name" : "detection_y0.jpg",
    "gpt_prediction" : "tumor",
    "gpt_raw_response" : "The image is an MRI scan of a human brain. There is a red b
}
```

Field descriptions:

- _id: Unique identifier for each document.
- image_name: The filename of the MRI image used for classification.
- gpt_prediction: The label assigned by GPT-4o to the highlighted area in the image (e.g., "tumor").
- gpt_raw_response: The complete text response provided by GPT-4o, which describes the image and the classification context.

Note: The structure above is provided as an example; actual values vary per document. Moreover, "the label says 'tumor'" is the answer obtained without any hint or indication in the prompt, but simply by loading the image with the bounding box.

A.2 BraTS19 Dataset Structure and Channel Mapping

In the BraTS19 dataset, each volume is multimodal and consists of 4 channels. The preprocessing code clearly shows the following correspondence:

```
channel_to_modality = {
    "channel_0": "FLAIR",
    "channel_1": "T1w",
    "channel_2": "T1gd",
    "channel_3": "T2"
}
```

In the `decathlon_dataset` collection construction code, each document contains:

```
collection: medical_images.decathlon_dataset
{
 "_id" : ObjectId("67dd72ab2dacbe1585625523"),
 "image_id" : 0,
 "image_shape" : [ 4, 240, 240, 155 ],
 "label_shape" : [ 3, 240, 240, 155 ],
 "slice_index" : 60,
 "image_channels" : {
  "channel_0" : "", "channel_1": "", "channel_2": "","channel_3": ""
 },
 "label_channels" : {
        "channel_0" : "", "channel_1": "", "channel_2": ""
    }
}
```

A.3 Structure of the `modality_predictions` Collection

The `modality_predictions` collection contains the results of modality prediction tasks, including both ground truth and model predictions for each slice.

Each document in the collection includes the following fields (the following JSON structure is provided as an example):

```
 collection:  medical_images.modality_predictions
{
  _id :  ObjectId("67e691522a74d9d17f59e287"),
  _source_doc_id :  "67dd72ab2dacbe1585625523",
  image_id :  0,
  slice_index :  60,
  channel :  "channel_0",
  true_modality :  "FLAIR",
  gpt_prediction :  "Unknown",
  gpt_raw_response :  "I can't identify the exact MRI modality type
  from the image."
}
```

Field descriptions:

- `_id`: Unique identifier for each document.
- `_source_doc_id`: Reference to the original document in the `decathlon_dataset` collection.
- `image_id`: ID of the image volume.
- `slice_index`: Index of the specific MRI slice.
- `channel`: The channel number (e.g., `channel_0` corresponds to FLAIR).
- `true_modality`: Ground truth MRI modality (e.g., FLAIR, T1w).
- `gpt_prediction`: The predicted modality by GPT.
- `gpt_raw_response`: The raw text response from GPT regarding the prediction.

Appendix B

Detailed Prompt Logs

This appendix contains detailed records of the prompt variations used with GPT-4o during the experimental phases. Each entry includes the prompt, a brief description of the image input, the ground truth classification, the model output, and whether the response aligned with the expected output.

Table B.1: Prompt variations and GPT-4o responses with example inputs from BR35H and BraTS19 datasets.

ID	Prompt Text	Input Image Description	Ground Truth	GPT-4o Output Summary	Correct?
P1	Describe the image.	BR35H MRI (no tumour)	No tumour	"Normal brain MRI with no abnormal findings visible."	Yes
P2	Describe the image.	BR35H MRI (tumour)	Tumour	"Irregular structure in the right hemisphere could suggest an abnormality."	Partial
P3	Describe the image and indicate whether a brain tumour is present or not.	BR35H MRI (tumour)	Tumour	"The image shows a hyperintense region consistent with a tumour."	Yes
P4	Describe the image and specify the quadrant(s) containing a tumour, if any.	BR35H MRI with quadrants (tumour in Q2)	Tumour in Q2	"The tumour appears to be located in quadrant 2, as indicated by abnormal tissue."	Yes
P5	Based only on the image, identify the MRI modality type. Choose from: FLAIR, T1w, T1gd, or T2.	BraTS19 slice (FLAIR)	FLAIR	"The high signal in peritumoural areas indicates this is likely a FLAIR sequence."	Yes
P6	[No prompt, raw image input]	BR35H MRI (tumour)	Tumour	"This is an image of a brain. No specific abnormalities detected."	No
P7	Describe the four-panel image: 1) MRI input, 2) Ground truth, 3) Model prediction, 4) Grad-CAM heatmap.	BraTS19 four-panel image	ET present	"Panel three shows predicted tumour area; panel four's heatmap overlaps well with ground truth."	Yes
P8	Does this image contain signs of a brain tumour? Answer: 'tumour', 'no tumour', or 'unknown'.	BR35H MRI (no tumour)	No tumour	"No tumour"	Yes

Legend:

- **Yes:** Output matched clinical ground truth.
- **Partial:** Ambiguous or incomplete, not clearly actionable.
- **No:** Incorrect or misleading output.

Appendix C

Code Snippets and Pseudocode

This appendix presents selected code snippets and pseudocode illustrating key components of the segmentation and explainability pipeline. The code is written in Python using PyTorch and MONAI libraries.

C.1 Extracting 2D Slices from BraTS Volumes

Script: `extract_slices_for_segmentation.py`

```
for patient in patients:
    load all 3D NIfTI files (T1, T1ce, T2, FLAIR)
    stack into a 4-channel volume: (4, H, W, D)
    for each axial slice (z):
        skip if label is empty
        resize image + label to (224, 224)
        save as .npz with keys: image, label
```

Example:

```
img_slice = img_volume[:, :, :, z]
label_slice = seg_volume[:, :, z]
np.savez_compressed(out_path, image=img_resized, label=label_resized)
```

C.2 U-Net Training with MONAI

Script: `train_segmentation_monai2d.py`

```
model = UNet(
    spatial_dims=2,
    in_channels=4,
    out_channels=4,
    channels=(32, 64, 128, 256),
    strides=(2, 2, 2),
    num_res_units=2,
    norm='batch'
)
```

Training Loop:

```
for epoch in range(n_epochs):
    for images, labels in train_loader:
        outputs = model(images)
        loss = DiceLoss(outputs, labels)
        optimizer.zero_grad()
        loss.backward()
        optimizer.step()
```

C.3 Generating Grad-CAM Heatmaps

Script: `gradcam_segmentation_interattivo.py`

```
model = load_trained_model()
cam_extractor = GradCAM(model, target_layer="model.2.1.conv.unit0.conv")

for image, label in dataloader:
    output = model(image)
    heatmap = cam_extractor(class_idx=3, scores=output)[0]
    visualize FLAIR + prediction + heatmap overlay
```

Overlay Example:

```
axs[3].imshow(flair, cmap="gray")
axs[3].imshow(heatmap, cmap="jet", alpha=0.5)
```

Save Interactive Output:

```
if user_input == 'y':
    fig.savefig("gradcam_{fname}.png")
```

C.4 Binarization of Grad-CAMs and IoU Computation

For quantitative evaluation of the Grad-CAM visualizations described in Section C.3, each heatmap was normalized to the [0,1] range on a per-image basis and then binarized using a fixed threshold of 0.5. This cutoff was selected following standard practice for gradient-weighted activation maps in medical imaging [39, 40] and was applied uniformly across all experiments, models, and segmentation classes (background, edema, enhancing tumor, NCR/NET). No class-specific or dataset-specific tuning was performed to ensure reproducibility.

The binarized activation maps were then compared with the predicted segmentation masks using Intersection over Union (IoU) and Dice coefficient. The following pseudocode summarizes the process used to generate the quantitative evaluation results:

```
for image in dataloader:
    output = model(image)
    for class_id in {0,1,2,3}:
        cam = GradCAM(output, target_class=class_id)
        heatmap_norm = (cam - cam.min()) / (cam.max() - cam.min())
        heatmap_binary = heatmap_norm > 0.5
        pred_mask = (argmax(output) == class_id)
        iou = compute_iou(heatmap_binary, pred_mask)
```

C.5 Dataset Loader for Training and Grad-CAM

Class: `Brats2DSegmentationDataset`

```
def __getitem__(self, idx):
    data = np.load(self.files[idx])
    image = torch.tensor(data["image"], dtype=torch.float32)
    label = torch.tensor(data["label"], dtype=torch.long)
    return image, label
```

Appendix D

GPT-4o Image Analysis Snippets

This appendix presents key pseudocode and representative code snippets used to analyze brain MRI slices using GPT-4o. The system encodes medical images as base64 and submits them to GPT-4o using structured prompts. Responses are parsed and stored in MongoDB for further analysis.

D.1 Image Encoding and Prompt Submission

Convert PNG to base64 and build OpenAI-compatible request:

```
def encode_image_base64(image_path):
    with Image.open(image_path) as img:
        buffer = BytesIO()
        img.save(buffer, format="PNG")
        img_b64 = base64.b64encode(buffer.getvalue()).decode("utf-8")
        return f"{img_b64}"
```

Send to GPT-4o:

```
response = openai.chat.completions.create(
    model="gpt-4o",
    messages=[{
        "role": "user",
        "content": [
            {"type": "text", "text": PROMPT},
            {"type": "image_url", "image_url": {"url": data_url}},
        ],
```

```
    }]
)
```

D.2 MongoDB Logging

Structure for logging GPT-4o output:

```
document = {
    "image_name": filename,
    "image_path": image_path,
    "gpt_description": response_text,
    "model_name": "gpt-4o",
    "processing_timestamp": datetime.utcnow().isoformat() + "Z"
}
collection.insert_one(document)
```

Quadrant-specific annotations (example vector [True, False, False, True]):

```
def extract_quadrants_vector(response_text):
    found = re.findall(r'\b[1-4]\b', response_text)
    return [i in set(map(int, found)) for i in range(1, 5)]
```

D.3 Robust Response Handling and Rate Limiting

Handle refusals and JSON parsing errors:

```
try:
    result_json = json.loads(response.choices[0].message.content)
except JSONDecodeError:
    result_json = {
        "has_tumour": None,
        "is_refusal": True,
        "refusal_reason": "Invalid JSON format"
    }
```

Auto-retry on rate limit:

```
def safe_gpt_call(*args, **kwargs):
    while True:
        try:
            return openai.chat.completions.create(*args, **kwargs)
        except OpenAIError as e:
            wait = parse_wait_time_from_error(e) or 10
            time.sleep(wait + 1)
```

D.4 Prompt Variations Used

Free-text prompt:

```
"Describe the image without assuming any labels or prior knowledge."
```

Classification prompt:

```
"Describe the image, indicating only whether there are tumours or not."
```

Quadrant prompt:

```
"Describe the image and indicate whether a brain tumor is present or not.
Indicate also and only the numbers of the quadrants where you see the tumor."
```

Appendix E

Graphical Interface for Explainable AI Results

This appendix presents the graphical user interface (GUI) developed for displaying segmentation results and associated explainability outputs in a clinical context. The interface was designed to facilitate clear and accessible presentation of AI-driven results using segmentation overlays, textual analysis from LLMs, and interactive components for user feedback.

E.1 Overview of Interface Workflow

The graphical interface supports the following sequential workflow:

1. **Image Selection:** Users upload or select an MRI image from the system.
2. **Start Segmentation:** Launches the automatic segmentation pipeline using a trained 2D U-Net model.
3. **XAI Tool Activation:** Triggers Grad-CAM generation and GPT-4o explanation rendering.
4. **Review Results:** Users view the segmentation mask, Grad-CAM heatmap, and textual explanation.
5. **Human Feedback Module:** The user is prompted to respond to a set of structured questions regarding three main aspects: the correctness of the segmentation output, the clarity and appropriateness of the explainability method (e.g., Grad-CAM), and the

coherence of the visual explanation provided by the language model. This guided evaluation aims to assess not only the system's output, but also the user's understanding and trust in the combined AI pipeline.

6. **Send Evaluation:** Submits structured feedback to the database for future system refinement.

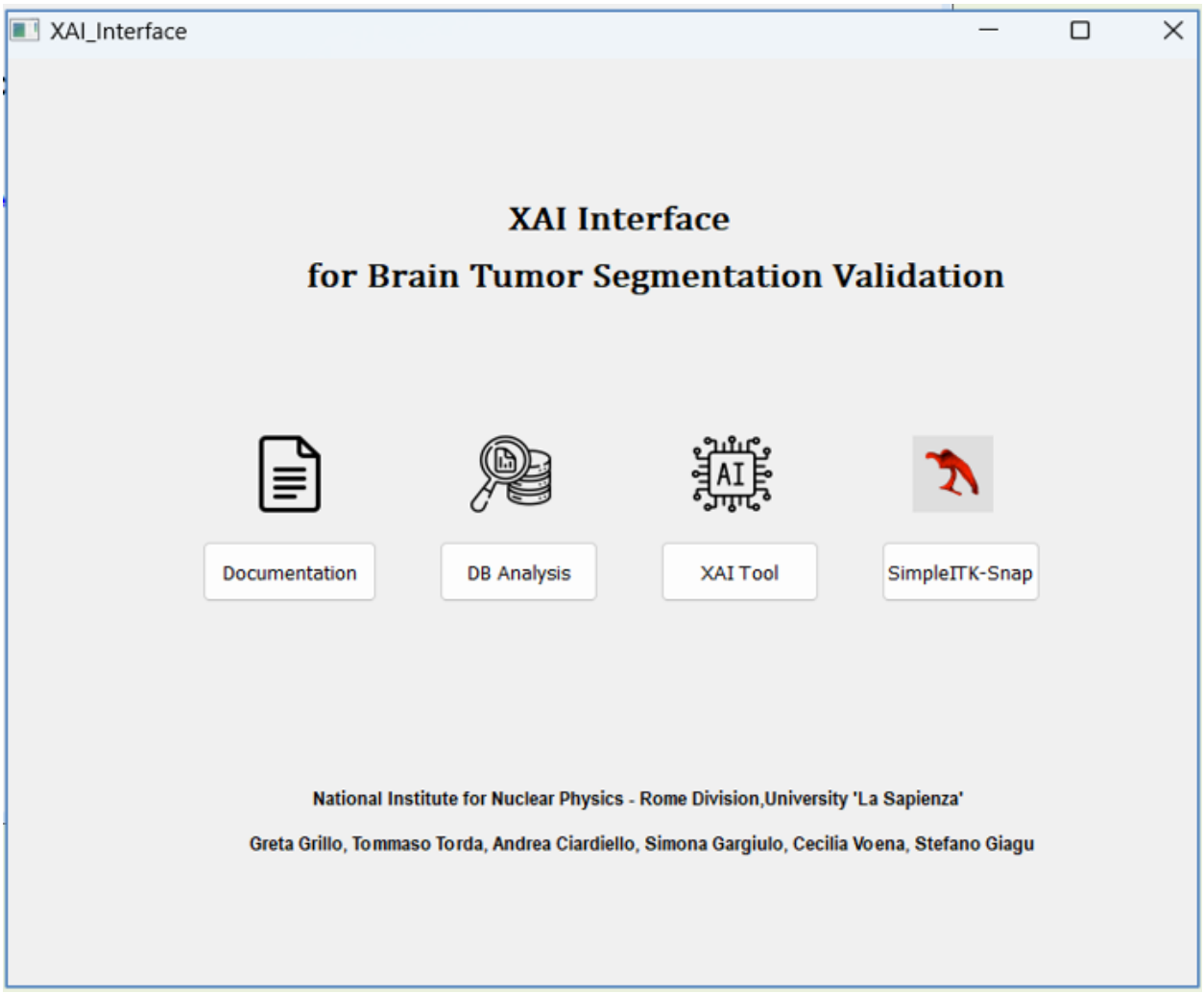


Figure E.1: Main window of the developed XAI Interface for Brain Tumor Segmentation Validation. The home screen provides access to the core modules: Documentation, Database Analysis, XAI Tool for segmentation and explainability, and external integration with SimpleITK-Snap for manual annotation and review.

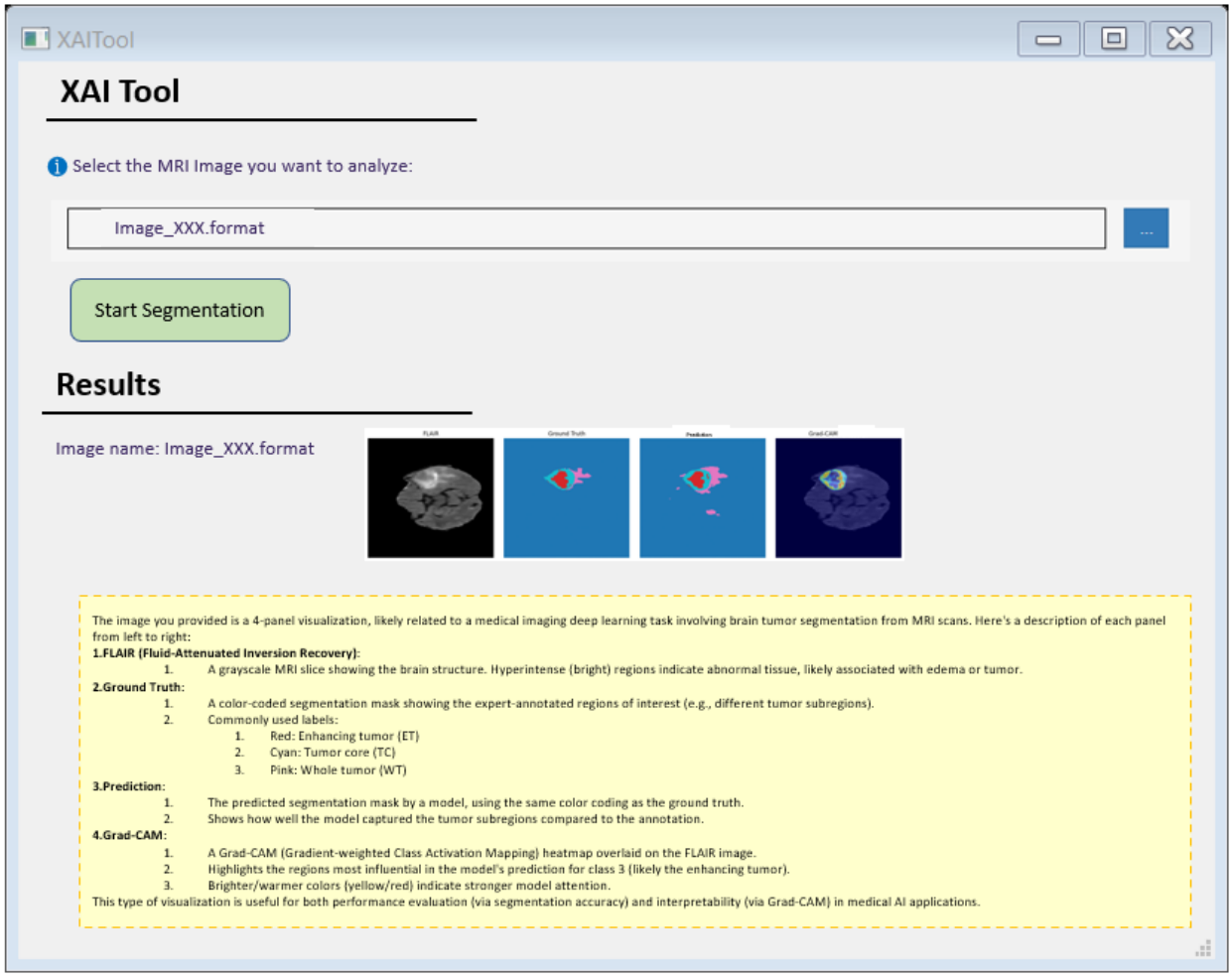


Figure E.2: Screenshot of the developed GUI (XAITool). The interface allows the user to load an MRI image, perform automatic segmentation, and visualise results through a four-panel view: original FLAIR slice (for example but not only FLAIR), ground truth, model prediction, and Grad-CAM heatmap. The colour-coded segmentation map and visual explanations support both clinical validation and interpretability of the AI pipeline.

E.2 GUI Components and Functionalities

- **MRI Image Display:** Visualizes the original slice alongside segmented and heatmapped outputs.
- **Explanation Panel:** Presents the textual description from GPT-4o summarizing tumor presence, location, and confidence.
- **Quadrant Annotation:** Displays labeled quadrants on the image to facilitate spatial feedback and verification.
- **Proponents/Opponents Tabs:** Allows two types of users to independently annotate the correctness or concerns regarding model outputs.
- **Feedback Checklists:** Users can select common error types such as:
  - Tumor incorrectly positioned
  - Labels swapped
  - Labeling incorrect within tumor area
- **Image Similarity Evaluation:** Additional panel to flag if control images are visually or semantically dissimilar to the target.

E.3 Principles Behind the GUI Design

The interface was designed around the following XAI usability principles:

- **Transparency:** Explicit representation of segmentation and LLM interpretation.
- **Completeness of Information:** Multimodal integration (image + text).
- **Accountability:** Collects user judgments on model decisions.
- **Adaptability to Knowledge Level:** Effective even for users with limited technical background.
- **Clear Format:** Organized panels to separate model output, visual explanation, and human review.