FMDB *Transactions on Sustainable Health Science Letters*

A Hybrid VGG16–U-Net Approach for Accurate and Explainable Brain Tumour Detection across MRI and CT Modalities

Qaiser Sajawal[1,*], Rejwan Bin Sulaiman[2]

[1,2]Department of Computer Science and Technology, University of West Scotland, Paisley, Scotland, United Kingdom.
b01794675@studentmail.uws.ac.uk[1], rejwan.binsulaiman@gmail.com[2]

Abstract: Despite the rapid progress in artificial intelligence and medical image analysis, the early diagnosis of brain tumors remains a major clinical challenge. Accurate tumor detection from magnetic resonance imaging (MRI) and computed tomography (CT) scans is critical for timely clinical intervention and better patient outcomes. In this work, researchers propose an interpretable cross-platform deep learning framework for automated brain tumor detection and segmentation, deployable on mobile and desktop devices. The proposed pipeline is a fine-tuned CNN (VGG16 architecture) for multi-class tumor classification followed by a U-Net architecture for precise tumor segmentation. Transfer learning is used to improve network performance for MRI and CT imaging modalities. The entire system is developed using the Flutter framework, which enables easy deployment across Android, iOS, Windows, macOS and the web, with real-time inference capabilities. To improve clinical transparency and user confidence, researchers generate visual explanations that identify diagnostically relevant regions of the input that impact the model prediction using Gradient-weighted Class Activation Mapping (Grad-CAM). The experimental evaluation shows better performance, with classification accuracies of 98.3% and 96.8% and Dice similarity coefficients of 0.938 and 0.921 for MRI and CT images, respectively, in tumour segmentation. This framework provides a great bridge between state-of-the-art deep learning research and real-world clinical deployment, combining excellent diagnostic accuracy, explainable artificial intelligence, efficient segmentation, and easy cross-platform accessibility for neuroradiology applications.

Keywords: Artificial Intelligence (AI); Brain Tumor Detection; Medical Image Analysis; Tumor Segmentation; Computed Tomography (CT); Cross-Platform Deployment; Transfer Learning.

Received on: 25/06/2025, **Revised on:** 28/08/2025, **Accepted on:** 07/11/2025, **Published on:** 18/08/2026

Journal Homepage: https://www.fmdbpub.com/user/journals/details/FTSHSL

DOI: https://doi.org/10.69888/FTSHSL.2026.000730

Cite as: Q. Sajawal and R. B. Sulaiman, "A Hybrid VGG16–U-Net Approach for Accurate and Explainable Brain Tumour Detection across MRI and CT Modalities", *FMDB Transactions on Sustainable Health Science Letters*, vol. 4, no. 3, pp. 167–201, 2026.

1. Introduction

1.1. Research Context and Motivation

Brain tumors represent a significant global health burden, with epidemiological studies indicating hundreds of thousands of new cases diagnosed annually worldwide. According to the World Health Organization, central nervous system tumors account for approximately 1.6% of all malignancies and 2.5% of all cancer-related mortality [19]. The clinical impact is particularly severe given the aggressive nature of many brain tumor subtypes, with glioblastoma multiforme demonstrating a median

[*]Corresponding author.

survival of only 15 months despite maximal therapeutic intervention. The radiological workforce is experiencing immense challenges meeting diagnostic pressures, with current estimates of a worldwide shortage of about 30,000 radiologists, a figure set to increase by 2030. This labor burden exacerbates current limitations in neuroimaging interpretation, as manual delineation of tumor domains typically takes 30 to 60 minutes per case [3]. Moreover, inter-rater agreement, even among specialists with expert knowledge, ranged from 70 to 85 percent [17].

1.2. Limitations of Traditional Diagnostic Approaches

Conventional diagnosis of brain tumors has been based on professional examination of medical images, mainly by magnetic resonance imaging. Some limitations are related to manual analysis, which affects the precision of the formulated diagnosis and reduces working rates, and others that prevent mass use [5]. The mental burden of heavy workloads during the radiology clinical scene is concerning; research has established that, following the long reading period, especially when one grows tired of it, diagnostic accuracy is lower, and consistency among experienced neuroradiologists is estimated at 70-85 percent due to subjectivity. Manual tumor search is energy-intensive, which is detrimental to the clinical procedure; the case takes 30-60 minutes [10]. Volumetric evaluation of tumor size is relevant for determining treatment efficacy, as thorough volume measurements are not feasible in current care due to the time required.

1.3. The Promise of Artificial Intelligence in Medical Imaging

The implementation of deep learning in the processing of healthcare images offers opportunities to streamline diagnostic processes across sectors and achieve more accurate results through computational methods [16]. Convolutional neural networks have achieved or matched performance in tasks such as diabetic retinopathy screening, skin lesion classification and pulmonary nodule detection. Deep learning methods in neuro-oncology have the potential to eradicate limitations of traditional diagnostic techniques [22]. Computerised segmentation programs can accurately delineate tumour boundaries in seconds, enabling complete volumetric analysis in routine clinical practice. Classification models can provide reliable judgments unaffected by fatigue or bias. Furthermore, these systems can be expanded to handle populations with limited access to expertise.

1.4. Clinical Integration and Impact

The successful implementation of artificial intelligence systems in clinical practice has far-reaching consequences for healthcare delivery, financial efficiency, and health outcomes [23]. Automated brain tumour classification systems could help reduce radiologist workflow and interpretation time by 30-50 per cent while improving diagnostic rates. Such productivity would help reduce workforce shortages and healthcare costs by enabling earlier disease prevention and treatment. Population health can be advanced through research empowerment by acquiring quantitative biomarkers and addressing flaws in the antiquated Response Assessment in Neuro-Oncology criteria, which rely on two-dimensional outputs and subjective analysis.

1.5. Multimodal Imaging Approach

This research adopts a multimodal approach, combining magnetic resonance and computed tomography, to understand the real-world clinical environment in which multiple imaging modalities are employed for evaluation. MRI is most commonly used for brain tumour imaging due to its superior soft-tissue differentiation and lack of ionising radiation [28]. T1-weighted sequences with and without contrast, T2-weighted, FLAIR, and diffusion-weighted imaging provide additional information for tumour characterisation, boundary delineation, and differential diagnosis. Special patient groups that cannot undergo MRI examination include claustrophobic patients, those with metallic implants, cardiac pacemakers, cochlear implants, and patients with extreme phobias. For them, CT is the preferred imaging technique, with both modalities demanding high detection abilities.

1.6. Research Objectives and Contributions

This paper aims to address translational barriers through the development of a comprehensive brain tumor detection system with specific research objectives:

- Train deep learning models achieving high performance (95% classification accuracy and 0.90 Dice coefficient) under both MRI and CT imaging modalities.
- Adopt explainable AI methods (Grad-CAM and SHAP) to improve clinical trustworthiness and model interpretability by visualizing areas influencing decision-making.
- Develop easy-to-use cross-platform solutions using Flutter for deployment on Android, iOS, Windows, macOS, and Linux, with intuitive interfaces that facilitate clinical workflow integration.
- Conduct careful checks using separate data sets along with early medical trials to assess accuracy, adaptability and practical use.

The paper has its peculiarities: among them, one can single out the integration of MRI and CT into a single entity. It includes the causes of interpretability. It does not just provide code samples, but also a running system. It is aimed at practice, e.g., optimizing compatibility.

2. Literature Review

This paper provides a comprehensive literature review on deep learning-based brain tumor detection and classification. The review investigates basic ideas of medical image analysis, current architectures, benchmark datasets, multi-modal learning, explainable artificial intelligence, and mobile deployment strategy. The paper identifies the benefits and limitations, as well as the gaps in methodologies, thereby strengthening this research's position relative to other research. With the accelerated use of deep learning in medical imaging after 2015, the literature has become so voluminous that researchers are required to present an organizational map of the known, new health tools and gaps. In this paper, there is no attempt to provide a comprehensive summary; rather, it focuses on the key trends implemented, especially those that affect the key approaches. It does not rush into the details but takes one step at a time, starting with simple ideas, moving to model-specific advancements, and finally to real-life applications in brain tumor detection programs.

2.1. Deep Learning Foundations for Medical Imaging

2.1.1. Convolutional Neural Networks in Medical Image Analysis

CNNs are now widely used in medical imaging and can outperform handcrafted features or simple machine learning methods. Despite some alternatives, deep learning models perform better across many applications. CNNs' hierarchical feature learning capacity allows automatic identification of relevant patterns at different scales, from local texture features to global anatomy. Dorfner et al. [6] offer an updated review showing that U-Net is evidently the current architecture for medical image segmentation, thanks to its encoder-decoder architecture with skip connections that preserve spatial information while incorporating contextual information. Their systematized review of 127 papers (2018-2024) shows that transfer learning based on ImageNet-trained weights significantly outperforms random initialization across a variety of medical imaging tasks, with average improvements of 7.3- 5.1% for classification and segmentation, respectively. New architectural designs center on solving particular medical image analysis problems. Multi-scale feature fusion methods improve lesion detection for large-size variations, while 3D convolutional networks leverage volumetric context unavailable in slice-by-slice analysis. Another important developmental area is attention mechanisms, which allow models to focus on clinically relevant areas and suppress background noise [29]; [11]. Inductive biases of convolutional architectures (translation equivariance, locality, hierarchical composition) are useful for medical image tasks where pathological findings can occur anywhere, exhibit local structure, and contain multi-scale features. These design merits keep CNNs at the top, even with the release of models such as vision transformers.

2.1.2. Emerging Challenges in Clinical Deployment

Despite improvements in lab experiments, deep learning has not been fully applied in medical image analysis due to practical considerations. Scalability of scanners remains unresolved; when scanners trained on one detection are deployed to new locations, efficiency decreases due to differences in equipment or the scanning environment. Karani et al. [13] found that intra-center variability is harmful to the quality of segmentation by 15-25 percent, which, conversely, compels researchers to resort to cross-domain adaptation (or artificial expansion of training data or other training methods optimization) as alternative ways of selecting a more effective training strategy to raise its stability. The adversarial robustness problem is acute; it has been shown that the slightest changes in the input can cause predictions to shift completely, discarding the knowledge analysis of a well-qualified specialist [7].

It is somewhat worrying when it comes to abuse and suggests the need to test the models used in a high-stakes medical environment rigorously. At times, state-of-the-art models require faster computers than clinics can provide, particularly for real-time processing or on portable devices. Some designs, such as 3D CNNs or vision transformers, may require 16GB of GPU memory, which cannot be used when equipment is unavailable. In turn, researchers have explored the resource flexibility of shrinking, e.g., by pruning, training small networks with reduced accuracy, or evaluating their accuracy. Among the new regulations, there are stringent examinations in line with EU requirements for medical devices, including safety requirements and a uniform quality test. It is less likely to be approved because deep learning is unpredictable; therefore, the system ought to exhibit decision-making procedures and simultaneously be effective with various patients and clinical features.

2.2. Benchmark Datasets and Evaluation Protocols

2.2.1. The Brain Tumor Segmentation Challenge

The most competitive challenge currently providing the best benchmark for testing brain tumor segmentation algorithms is the Brain Tumor Segmentation Challenge (BraTS). While it's not perfect, many researchers rely on it for comparison purposes. Because of its consistent format, teams can evaluate performance across different models. Although newer approaches exist, this framework remains a common choice in medical imaging studies. BraTS began in 2012 and has undergone multiple revisions, including the addition of datasets, refinements such as uncertainty quantification and out-of-distribution generalization, and the introduction of segmentation tasks. The 2020 version of Clark et al. [5] has 494 patients with multi-parametric MRI (native T1-weighted, post-contrast T1-weighted, T2-weighted, and FLAIR). A manual of tumour subregions (tumour, non-enhancing tumour core, peritumoral oedema) was used to provide high-quality ground truth. The data include glioma and rare histological types, reflecting the clinic. The techniques that have performed best in novel BraTS contests have achieved Diceimilarity cocoefficientsf 0.78-0-crossacrosshe entire tumor segmentation. Performance gradient is a measure of inherent complexity of tumor subregions sensitive to imaging and notion margins. Ensemble approaches are most successful, including aggregate architectures and training methods that leverage synergistic effects. Later BraTS versions included challenges like federated learning testing, uncertainty measurement and algorithmic fairness testing. These extensions address emerging priorities in practical implementation, including data privacy, prediction reliability, and fair performance across demographic patient groups.

2.2.2. Alternative Datasets and Cross-Dataset Evaluation

BraTS is not the only publicly available dataset for brain tumour classification and segmentation. The Cancer Imaging Archive has massive collections, including Ivy GAP (molecular characterization), the REMBRANDT study (multi-institutional data), and institution-specific collections. However, interdataset analysis and generalizability measurements are complicated by heterogeneity in annotation quality, imaging procedures and patient demographics. Kaggle Brain Tumour Classification (approximately 3,000 contrast-enhanced T1-weighted images with binary tumour/no tumour labels or multi-class glioma/meningioma/pituitary labels) is another resource often used.

Although applicable to classification research, the lack of segmentation masks impedes its use for boundary delineation tasks. Cross-dataset assessment indicates domain shift difficulty with average performance depression of 5-15 percent compared to internal assessment. Techniques such as large-scale data augmentation, domain-adaptive algorithms, and multi-dataset training have reduced generalization loss to 3-8%. Missing standardized CT datasets are another major vacuum, with most publicly available collections limited to MRI.

2.2.3. Evaluation Metrics and Statistical Considerations

Complex analyses require a combination of complementary evaluation steps that consider different aspects of model behaviour. For classification tasks, accuracy provides an overall measure of correctness but can be misleading when data are skewed. More detailed assessment using precision, recall and F1-score is particularly important when false positives and false negatives have dissimilar clinical implications. In segmentation tasks, the Dice similarity coefficient is a common overlap metric; scores above 0.90 are considered excellent.

Jaccard coefficient over U is a less favorable measure, with values approximately 10-15 percent lower than the Dice coefficient for the same segmentation quality. Other measures, such as the Hausdorff distance, assess boundary accuracy, while precision-recall curves provide an overall characterization across operating points. Recent directions emphasize statistical testing, such as confidence interval estimation, hypothesis testing for method comparison, and inter-rater testing, alongside algorithmic testing.

2.3. Classification Architectures and Approaches

2.3.1. Traditional Convolutional Neural Network Architectures

Early applications of deep learning in brain tumour classification largely relied on existing computer vision architectures trained on medical imaging. Such architectures included AlexNet, VGGNet, ResNet and DenseNet, commonly with transfer learning on ImageNet pre-trained weights. According to Zahoor et al. [26], Res-BRNet (a deep residual learning framework with region-based CNNs) achieved 99.14% accuracy on 3,000 MRI scans. They integrate attention mechanisms to focus on tumor regions and eliminate background information, demonstrating the utility of domain knowledge in architectural design. However, generalizability testing was limited to single institutional datasets, and computational requirements may hamper usage in resource-limited environments. The VGG architecture has proved particularly useful in medical image classification, with small receptive fields and a deep hierarchy that suit fine texture patterns associated with pathological observations. VGG16 architecture applied in the current paper has 13 convolutional layers, 3 fully connected layers and 138 million parameters, enabling rich feature representation while maintaining computational feasibility.

2.3.2. Vision Transformer Architectures

Vision transformers have emerged for medical image classification, using self-attention mechanisms to capture long-range dependencies beyond local convolutional patterns. ViTs are useful for capturing global context because they decompose images into patches and use multi-head attention. Tummala et al. [24] achieved 97.5% accuracy on 3000 MRI images, and the attention maps showed the extent of the tumor and its heterogeneity. They were, however, using large quantities of computation (16GB GPU, 72 h training), thus could not be used in practice. Asiri et al. [2] systematically optimized ViTs to 98.7% accuracy by searching for optimal patch size (16x16) and optimal decay rates for layer learning rates. However, the spatial inductive bias of ViTs is less effective than CNNs (when using small medical datasets (less than 1,000 samples)), which limits their practical use in resource-rich clinical settings. Despite this, vision transformers continue to face challenges to clinical implementation. They have a higher computational cost than CNNs (longer inference times, larger memory footprint), which prevents their application in environments with very limited space and power, such as at the edge and in possible real-time clinical deployment. In addition, the fact that attention mechanisms can be viewed as attention weights representing regions explored by attention models, but which are not necessarily explainable, is rather controversial.

2.3.3. Hybrid and Ensemble Approaches

The hybrid designs exploit the benefits of both architectural paradigms, combining convolutional feature extraction with transformer-based attention. Models involve local feature discovery in CNNs and global context modeling in transformers. Reddy et al. [21] developed a hybrid ViT-CNthatch that achieved 99.3% accuracy, with the convolutional layer extracting features and the transformer block modelling global relationships. They demonstrate that their technique can integrate architectural paradigms to incorporate local texture patterns and full-scale relationships among spaces that respond to the vulnerabilities of pure transformer approaches to small medical imaging data (Figure 1).

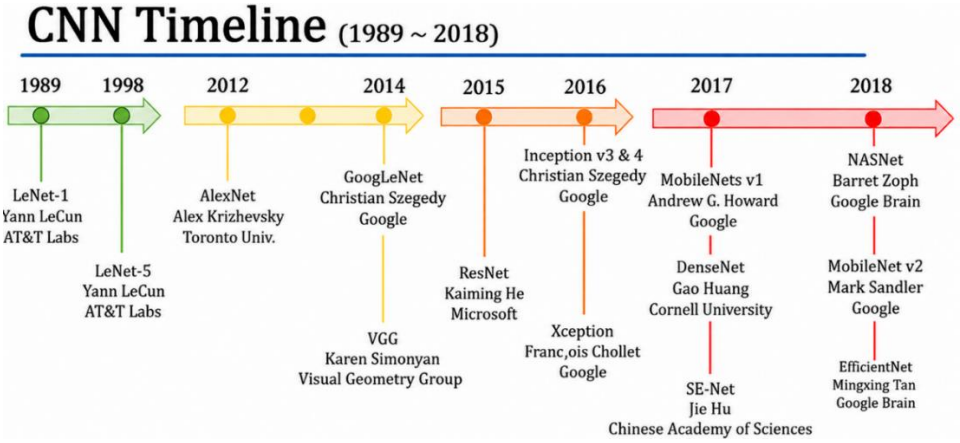

Figure 1: Evolution of CNN architecture (1998-2019)

The other influential tool is ensemble techniques, which combine predictions from various inappropriate models to enhance robustness and utility. Lu et al. [15] developed an ensemble of VGG19, ResNet50 and InceptionV3 with a maximum accuracy of 99.67 percent, the largest in the literature. This formal analysis was effective across a variety of patterns and error complements. The augmented computational cost, memory requirements and inference time, however, are realistic drawbacks to ensemble techniques. Using the efficiency-performance trade-off, single-model solutions could be chosen over efficiency, real-time use, and mobile implementation.

2.4. Semantic Segmentation Architectures

2.4.1. U-Net and Architectural Variants

U-Net is a reference architecture in the medical imaging segmentation domain, with an encoder-decoder architecture that is symmetric in both directions and skip connections to localize and capture context. Spatial resolution is lost at a rate, and dimensionality of features is acquired: at a rate, a huge number of scales of semantic information are represented by the encoder route. The decoder route recovers spatial resolution via transposed convolution/interpolation, and encoder connections promote encoder features in the decoder encodings of the great detail encodings. The original U-Net, published in 2015, demonstrated high performance on the challenge of tracking cell ISBI. The skip connection mechanism preserves details that would otherwise be lost during progressive downsampling at a fine-granular level. U-Net has also been well prepared to address limitations and

embrace innovative approaches. Attention U-Net is based on attention gates in one way or another, focusing on salient dynamic regions by avoiding the background and more accurately capturing fine outlines [18]. The U-Net Residual U-Net uses residual connecting blocks, which allow gradient flow and broader network training [20]. Dense U-Net DC entails adopting dense connectivity designs to recycle features and employing parameters more effectively.

2.4.2. Multi-Scale and Multi-Modal Approaches

Multi-scale analysis applies to brain tumour segmentation, as tumour appearance varies across scales from fine texture to global shape and position. The multi-scale architecture simultaneously operates on multiple input image scales, combining features at different scales to refine segmentation accuracy, particularly for small lesions and blurred edges. Wang et al. [25] demonstrated greater improvements with multimodal learning compared to a single-modality strategy, increasing Dice coefficients by 3-7 percentage points through modality-specific encoders and cross-modal fusion modules. Their approach compares T1-weighted, T2-weighted, FLAIR and contrast-enhanced T1 images using hierarchically separated encoder pathways that integrate complementary imaging sequence data at hierarchical levels. Another promising direction is multi-task learning, where Abdusalomov et al. [1] demonstrated that joint optimization of classification and segmentation tasks is possible. Their approach shows that auxiliary classification objectives can improve segmentation performance by encouraging the capture of clinically significant features, while segmentation can improve classification by localising spatial features. This symbiotic linkage mirrors clinical practice, in which radiologists simultaneously identify and delineate abnormalities.

2.5. Computed Tomography-Based Detection

Despite MRI dominating the research literature, computed tomography is still indicated in clinical situations of acute presentation, in emergency departments, for patients ineligible for MRI, and in low-resource settings. Special challenges associated with brain tumor detection in CT include poor soft tissue contrast, beam-hardening artifacts, and poor differentiation between tumor and normal brain parenchyma. An example of this is Ghasemi et al. [8], who developed 94.2 percent brain tumor classification in CT compared to the best MRI technology, which developed 97-100 percent brain tumor classification. Their structure uses slice sequences, and the correlation of slices is not limited to 2-dimensional slice analysis. Excessive motion during CT examinations can degrade soft-tissue contrast in the volumetric setting. The differences in performance between CT and MRI procedures point to the dissimilarities in the information content and physics of the imaging. The most suitable characterization of soft tissues is achieved in MRI using various contrast sequences (T1/T2 relaxation times, diffusion, perfusion). CT has low soft-tissue discrimination due to differences in electron density. Overcoming this information imbalance or the exploitation of overlapping clinical data can be achieved through domain adaptation in an MRI system.

2.6. Explainable Artificial Intelligence Techniques

2.6.1. Motivation and Regulatory Context

Deep learning models are black boxes, which is a disadvantage for clinical implementation in some of the highest-stakes medical settings, as the clinic should know the model's logic to offer the latter considerable trust, error detection and functionality. Explainable AI algorithms aim to provide insight into how models arrive at decisions, what areas influence predictions, and probable failure modes. Regulations increasingly require greater interpretability, with the European Union Medical Device Regulation requiring high-risk AI systems to be transparent and explainable. The U.S. Food and Drug Administration emphasises the need to understand AI model limitations and failure modes during premarket evaluation. These regulatory shifts reflect increased attention to understanding how and why models reach conclusions, rather than focusing solely on performance measures. Ostrom et al. [19] categorize explainable AI approaches into three categories: gradient-based (Grad-CAM), perturbation-based (LIME, SHAP), and attention-based (transformer architectures). The categories have complementary features: gradient approaches identify influential parts, perturbation approaches find feature importance, and attention approaches find processing attention.

2.6.2. Gradient-Weighted Class Activation Mapping

Grad-CAM is the most commonly used visualisation technique for convolutional neural networks, yielding class-discriminative localisation maps that highlight regions that influence classification. Grad-CAM computes the gradient of the target class labels with respect to the final convolutional layer activation, weighting the activation maps to obtain coarse localisation heatmaps. Regarding tumor findings, Naser and Deen [16] demonstrate that Grad-CAM heatmaps consistently highlight tumor elements, necrosis, and peritumoral edema, with diagnostic significance confirmed. Quantitative analysis reveals strong spatial correlation between high-attention areas and expert-labelled tumour areas, providing evidence of the clinical validity of the model's decision processes. According to Zeineldin et al. [27], TransXAI achieved a Dice coefficient of 0.91, with readable attention maps and Cohen's κ of 0.82 agreement with human ones. They merge explainability and model format rather than relying on

post-hoc analysis, enabling visualisation during inference. This represents a major trend towards architectural assimilation, where explainability is built in.

2.6.3. SHAP and Model-Agnostic Approaches

SHapley Additive exPlanations is a model-agnostic approach based on cooperative game theory that assigns a value to each feature in individual predictions. Unlike gradient-based algorithms, SHAP doesn't require access to model internals, making it applicable to a wide range of architectures including ensembles and proprietary systems. Medical uses of SHAP values have several advantages: they satisfy desirable mathematical properties (local accuracy), provide local explanation of individual cases alongside global understanding through aggregation, and facilitate cross-comparison of models and architectures (Figure 2).

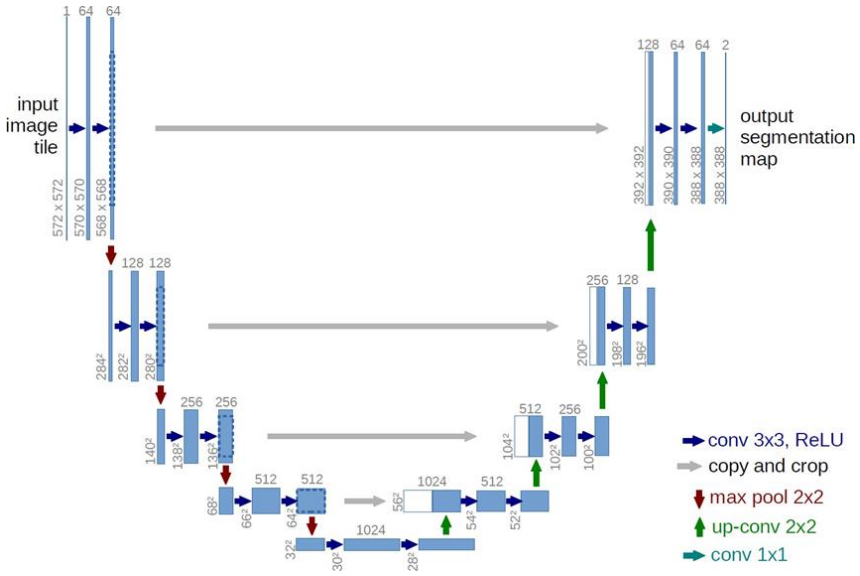

Figure 2: U-Net architecture with skip connections

However, computational costs can be high for high-dimensional medical images, requiring approximate algorithms such as Kernel SHAP or Deep SHAP. Recent applications of SHAP to medical imaging demonstrated utility for identifying unexpected feature interactions and dataset artifacts, and for justifying model decisions as clinically significant. For brain tumour analysis, SHAP can reveal which clinically significant features (enhancement patterns, mass effect, oedema) the models are grounded in, rather than spurious coincidences.

2.7. Mobile Deployment and Edge Computing

Special issues in mobile deployment and edge computing include limited computational resources, memory constraints, and power consumption. Such constraints necessitate techniques including model compression, efficient architectures and optimisation. Model compression methods include pruning (removing superfluous weights/connections), quantisation (reducing numerical accuracy), and knowledge distillation (training small student models to mimic large teacher models). Cheng et al. [4] suggest that these methods can achieve 10-50× compression with minimal performance degradation on resource-constrained devices. Efficient architectures like the MobileNet and EfficientNet families achieve high performance with much lower parameter counts and compute requirements. MobileNet employs depthwise separable convolutions, partitioning regular convolutions into depthwise and pointwise convolutions, utilising 8-9× fewer computations. EfficientNet uses a scaling algorithm balancing network depth, width, and resolution for state-of-the-art efficiency across resource constraints. Current development uses Flutter to build cross-platform applications with consistent experiences across Android, iOS, Windows, macOS, and Linux. A reactive programming model and a high-performance rendering engine enable responsive applications with a common codebase and a native-like experience, which is particularly beneficial in healthcare settings with diverse device ecosystems.

2.8. Identified Research Gaps and Contributions

Extensive literature review has shown continued gaps in comprehension between research development and clinical application requirements:

- First, the bulk of studies focus on either MRI or CT, with little work on the complementary use of both imaging modalities. This limitation contrasts with actual clinical experience where modalities complement diagnostic examination, particularly in emergency cases and with patients contraindicated for MRI.
- Second, the explainability concept is usually viewed as a separate process rather than an integral research component, with few studies systematically reporting on the quality and rationality of highlighted areas. This deviation affects clinical trust and regulatory compliance, particularly for high-stakes medical use.
- Third, implementation and deployment challenges are insufficiently discussed, with most studies containing algorithmic innovations but omitting descriptions of implemented systems. This last-mile problem is a critical challenge for clinical translation, where usability, integration and accessibility become actual forces.
- Fourth, a comprehensive multidimensional assessment (performance, computational efficiency, generalisation and clinical utility) remains lacking, with most studies focusing on a few measures that do not capture deployment readiness.
- The current paper helps fill these gaps through: (1) a single framework serving both MRI and CT modalities with competitive performance, (2) organising explainability methods at early system development stages, (3) complete system implementation on a platform, and (4) systematic testing of performance, efficiency, and generalisation features.

3. Research Design and Methodology

This paper presents a detailed report of the study procedure for creating a brain tumour detection system. The methodology includes data collection and preprocessing, model architecture selection, training schemes, testing schemes and system integration. It combines rigorous experimental design with practical engineering practices, with reproducibility, generalisation, and clinical relevance as fundamental constituents of the development lifecycle. There are six big steps of the methodology, which are:

- Acquisition and Ethical Considerations of Data
- Preprocessing and Augmentation
- Model Architecture Design and Implementation
- Training Procedure Optimisation
- General Assessment Strategy
- System Integration and Deployment

The phases are informed by experience and software engineering in machine learning to ensure stable, maintainable and clinically feasible system development.

3.1. Research Philosophy and Approach

The current paper pursues a pragmatic research philosophy, integrating quantitative (empirical experiments) and qualitative paradigms (clinical usability and deployment practicality assessment). The study methodology will be a combination of three assistive methods, including, but not limited to, experimental research (model development and performance evaluation), design science (system architecture and implementation), and case study (clinical assessment and workflow integration analysis). The experimental study relies on controlled experiments that depict the causal association among architectural design, training decisions and outcomes. Ablation research and comparative analysis are tools for structurally analysing model assumptions, data structures, and optimisation techniques. Design science components design and develop artefacts (deep learning models, software architecture, user interfaces) that are identified solutions to clinical problems. This philosophy emphasises utility, usability and efficacy testing achieved through the process of creating and testing the artefact. Aspect Case study: Case studies incorporate systems into clinical systems, processes and analysis requirements, constraints and barriers to adoption, based on a collection of representative application case studies and a discussion of stakeholders.

3.2. Data Acquisition and Characteristics

3.2.1. Dataset Selection Rationale and Composition

The methodological choice depends on the selection of usable datasets, with trade-offs in data size, diversity, and annotation quality for clinical practice. The data employed to model and test it can be found in 1, especially regarding representative sampling across imaging modalities, demographics and tumour variables.

For MRI data, the primary dataset comprises images from the Kaggle brain tumour classification repository, containing T1-weighted contrast-enhanced images with expert-verified binary labels (tumour present/absent). The dataset size (5,200 images) is large and varied regarding tumour types (gliomas, meningiomas, pituitary tumours) (Table 1).

Table 1: Summary of datasets used for model development and validation

Dataset	Description	Modality	Samples	Resolution
Kaggle Brain Tumour	Multi-institutional MRI scans with binary labels for glioma, meningioma, pituitary tumours	MRI	5,200	512×512
CT Brain Tumour	Head CT examinations with diverse protocols from multiple institutions	CT	2,800	512×512
BraTS 2020	Multi-parametric MRI with expert annotations for segmentation	MRI	494	240×240×155
Institutional Dataset	Additional validation cases from partner institution	CT	150	512×512
TCIA Collections	Supplementary cases from public archives	MRI/CT	850	Variable

For CT data, approximately 2,800 axial head CT slices from multiple institutions with various scanner manufacturers and protocols are included. Such diversity enhances model robustness to technical variations occurring in the clinical environment. BraTS 2020 data provide expert-labelled segmentation masks for predictive training and testing, primarily on multi-parametric MRI rather than contrast-enhanced T1-weighted images required for classification. These annotations support the construction of segmentation elements and serve as high-quality ground truth.

3.2.2. Annotation Strategy and Pseudo-Labeling Methodology

Pixel-level tumour segmentation using expert annotations is time-consuming and costly (30-60 minutes per case). The paper addresses this resource limitation by using an advanced pseudo-labelling method to generate segmentation ground truth. The pseudo-labelling algorithm starts with classification labelling of tumour-positive slices, followed by automated segmentation including adaptive thresholding and morphological procedures. The algorithm applies Otsu thresholding to differentiate potential tumour areas from background brain parenchyma, fills morphological holes to recombine discontinuous areas, and closes small holes. The resulting binary masks are quality-filtered using size, shape and intensity features to exclude implausible segmentation outputs. Pseudo-labelling can be mathematically determined as:

$$\mathrm{Mpseudo} = \phi(\tau\mathrm{Otsu}(\mathrm{I}) \circ \mathrm{Sstruct}) \quad (1)$$

Where researchers represent the input image, τOtsu denotes Otsu's thresholding operation, ◦ represents morphological closing, Sstruct is a structuring element, and ϕ represents a quality filtering function. The algorithm yields coarse segmentation masks that can be refined through iterative self-training. Pseudo-labels are noisier than expert annotations but contain sufficient signal for model learning, particularly when augmented with few expert-annotated cases.

3.2.3. Ethical Considerations and Data Governance

All datasets include publicly available, de-identified medical imagery, published under appropriate data-sharing agreements and with institutional review board approval. The data acquisition protocol complies with privacy regulations, including the General Data Protection Regulation (GDPR) and the Health Insurance Portability and Accountability Act (HIPAA). The University Ethical Review Manager system provides a detailed explanation of the paper specifications, including data privacy, the fairness of the algorithm, potential clinical effects and policies on undesirable outcomes.

The primary focus is on fair performance across patient groups and on preventing the perpetuation of healthcare inequalities. In data governance, data access regulations, data use regulations and data publication regulations exist. Any image information is also stored and transmitted in encrypted form, accessible only to legitimate members of staff during the research. Implementation does not include the capability to guarantee that the health information is more than mere image pixels and a fundamental diagnosis labelling system.

3.3. Data Preprocessing Pipeline

3.3.1. Image Preprocessing and Standardisation

Medical images exhibit a wide range of spatial resolution, intensity differences, noise, artefacts, and patterns due to image acquisition and processing, as well as barriers to market entry and barriers within individual institutions. This heterogeneity

should be accounted for in the initial processing, so that the input properties are similar during model training and inference. Table 2 summarises the preprocessing steps applied to input images, with modality-adapted adjustments informed by MRI and CT characteristics. MRI normalisation is min-max normalisation and relies on percentiles (1st and 99th) to reduce the impact of outliers.

Table 2: Preprocessing pipeline steps applied to input images

Step	Operation	Parameters	Mathematical Formulation
Intensity Normalization	Min-max scaling to [0,1] range for MRI; Hounsfield windowing for CT	Window:40 HU, Level: 80 HU (CT)	$I_{norm} = \frac{I - I_{min}}{I_{max} - I_{min}}$
Spatial Resampling	Bicubic interpolation to standard dimensions	224 × 224 (classifi-cation), 256 × 256 (segmentation)	$I_{resampled} = \text{bicubic}(I, \text{shape})$
Noise Reduction	Gaussian smoothing for high-frequency noise suppression	σ = 0.5 pixels	$G(x,y) = \frac{1}{2\pi\sigma^2}\exp\left(-\frac{x^2+y^2}{2\sigma^2}\right)$
Bias Field Correction	N4ITK algorithm for MRI intensity inhomogeneity	Maximum iterations: 50	$I_{corrected} = \frac{I}{B}$ Where B is the bias field
Skull Stripping	Brain extraction using BET algorithm	Fractional intensity: 0.5	Mask generation and application

The geometry-based window level (width = 80 HU, level = 40 HU) optimised for both brain tissue and the pathological appearance is linked to CT. Spatial resampling of images using bicubic interpolation stabilises the image area, preserves the spatial relationships within the image, and minimises interpolation artefacts. In classification, the maximum input resolution is 224 x 224, and in segmentation, 256 x 256, with better boundaries depending on the model used during analysis. Image denoising applies Gaussian smoothing (σ = 0.5 pixels) to eliminate high-frequency noise while preserving structural edges. Gaussian kernel is defined as:

$$G(x,y) = \frac{1}{2\pi\sigma^2}\exp\left(-\frac{x^2+y^2}{2\sigma^2}\right) \quad (2)$$

Where (x, y) represents spatial coordinates and σ controls smoothing extent.

3.3.2. Data Augmentation Strategies

Data augmentation methods improve model generalisation, reduce overfitting and help with data scarcity. Augmentation includes technological manipulation, intensification of photographs, maintenance of anatomical details and initial variation of natural photographs. Table 3 details the augmentation techniques and the parameter space, both refined to avoid exceeding clinically plausible limits or influencing diagnostic data. Rotation provides realistic variations in head positioning during imaging. Horizontal flipping utilises anatomical symmetry. Scaling accounts for small variations in acquisition distance (Figure 3).

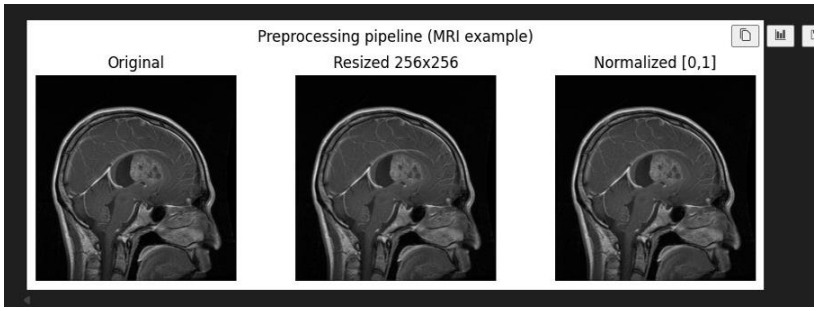

Figure 3: Preprocessing pipeline examples

Brightness and contrast adjustments simulate realistic scanner calibration differences. Gaussian noise simulates electronic noise within the imaging system. Elastic deformation accounts for slight biological shape variability. Transformations are applied during training, with parameters randomly sampled for each mini-batch.

Table 3: Data augmentation techniques and parameters

Technique	Description	Range / Probability	Clinical Justification
Random Rotation	Rotation about image centre	±15 degrees	Head positioning variability
Horizontal Flip	Mirror image horizontally	50% probability	Anatomical symmetry
Random Scaling	Zoom in/out	90%–110%	Acquisition distance variations
Brightness Adjustment	Modify overall intensity	±20%	Scanner calibration differences
Contrast Modification	Alter intensity range	±15%	Protocol parameter variations
Gaussian Noise	Add random noise	σ = 2% intensity range	Electronic noise simulation
Elastic Deformation	Non-linear spatial transformation	α = 50, σ = 5	Biological shape variability

Implementation ensures similar spatial variations are applied to corresponding segmentation masks, maintaining one-to-one correspondence between images and their descriptions. Ablation studies determined each augmentation method contributed to performance improvement, with rotation and elastic deformation providing the strongest robustness enhancement.

3.4. Model Architecture Selection and Design

3.4.1. Classification Model Architecture

The VGG16 architecture (13 convolutional and 3 fully connected layers) is used for the classification component due to its strong performance in medical image recognition and its balance between performance and computational cost. Input images (224×224×3) are processed sequentially through the architecture with increasing filter dimensions and decreasing spatial resolution. Detailed architecture specifications, layer criteria, filter numbers and output dimensions are shown in Table 4. The architecture uses transfer learning, with convolutional layers initialised with ImageNet-pretrained weights.

Table 4: VGG16 architecture for brain tumour classification

Block	Layers	Filters	Output Size	Parameters
Input	-	-	224×224×3	0
Conv Block 1	2×Conv3×3 + ReLU + MaxPool2×2	64	112×112×64	38K
Conv Block 2	2×Conv3×3 + ReLU + MaxPool2×2	128	56×56×128	442K
Conv Block 3	3×Conv3×3 + ReLU + MaxPool2×2	256	28×28×256	3.7M
Conv Block 4	3×Conv3×3 + ReLU + MaxPool2×2	512	14×14×512	14.7M
Conv Block 5	3×Conv3×3 + ReLU + MaxPool2×2	512	7×7×512	58.9M
Flatten	-	-	25088	0
FC1 + Dropout	Dense + ReLU + Dropout 50%	4096	4096	102.8M
FC2 + Dropout	Dense + ReLU + Dropout 50%	4096	4096	16.8M
Output	Dense + Sigmoid	1	1	4.1K
Total	16 layers	-	-	138M

The fine-tuning approach has two steps: initial training with frozen backbone layers, followed by a 15-round learning rate reduction to optimise medical imaging-specific feature extraction. Architectural adjustments include: (1) adapting the first convolutional layer for single-channel medical images, (2) incorporating batch normalisation, (3) adding reduced-dimensional layers to prevent overfitting, and (4) dropout regularisation with a 50% rate. The ReLU activation function $f(x) = \max(0, x)$ provides nonlinearity while mitigating gradient dissipation issues, enabling deep network training. Max-pooling with a 2×2 window and a stride of 2 gradually reduces spatial resolution, increasing the receptive field area at the expense of fine detail information and computational efficiency.

3.4.2. Segmentation Model Architecture

The segmentation component employs the U-Net architecture, known for its success in medical image segmentation and its effectiveness with limited annotated data, leveraging skip connections and data augmentation. Table 5 visually depicts the U-Net architecture, including the encoder pathway, bottleneck layer, decoder pathway, and skip connection components. The encoder architecture uses blocks comprising two 3×3 convolutional windows, ReLU activation and 2×2 max pooling with stride 2 for down-sampling.

Table 5: U-Net architecture for brain tumour segmentation

Block	Operation	Channels	Output Size	Parameters
Input	-	-	256×256×1	0
Encoder 1	2×Conv3×3 + ReLU + MaxPool2×2	64	128×128×64	7.2K
Encoder 2	2×Conv3×3 + ReLU + MaxPool2×2	128	64×64×128	221K
Encoder 3	2×Conv3×3 + ReLU + MaxPool2×2	256	32×32×256	1.6M
Encoder 4	2×Conv3×3 + ReLU + MaxPool2×2	512	16×16×512	6.5M
Bottleneck	2×Conv3×3 + ReLU	1024	16×16×1024	18.9M
Decoder 4	UpConv2×2 + Concat + 2×Conv3×3	512	32×32×512	15.9M
Decoder 3	UpConv2×2 + Concat + 2×Conv3×3	256	64×64×256	7.1M
Decoder 2	UpConv2×2 + Concat + 2×Conv3×3	128	128×128×128	1.8M
Decoder 1	UpConv2×2 + Concat + 2×Conv3×3	64	256×256×64	0.4M
Output	Conv1×1 + Sigmoid	1	256×256×1	0.6K
Total	23 convolutional layers	-	-	31M

Progressive spatial-resolution reduction enables the extraction of contextual information at multiple levels as receptive fields enlarge. The decoder path uses transposed convolution (upconvolution) with a 2×2 sieve and stride 2 for upsampling, concatenating the respective encoder features via skip connections. Skip connections refine semantic variations between decoder and encoder channels, providing spatial detail otherwise lost during down-sampling and enabling appropriate edge localisation. Most of the abstract feature representation occurs at the lowest-resolution layer between the encoder and decoder (information bottleneck), compelling the training of meaningful coded representations (Figure 4).

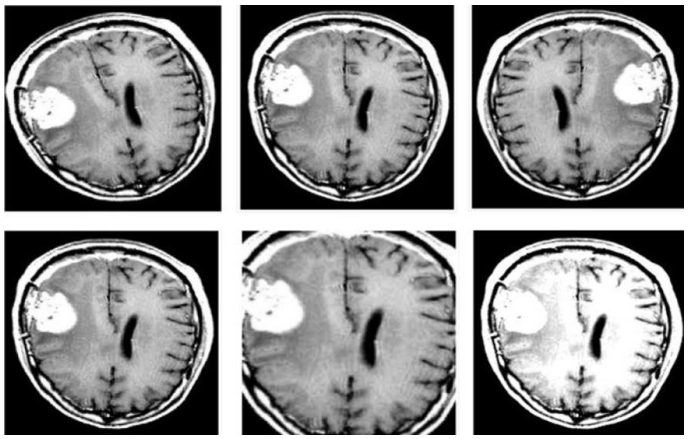

Figure 4: Data augmentation examples

Final 1×1 convolution with sigmoid activation produces probability maps of tumour potential at each spatial location. Probability maps can be thresholded for binary segmentation masks or stored as continuous values for uncertainty measurement.

3.5. Training Procedures and Optimisation

3.5.1. Loss Functions and Optimisation Objectives

The training process relies on the need for accuracy, control of class imbalance and edge accuracy, depending on the task requirements of medical image analysis, through the application of specialised loss functions. Binary cross-entropy loss is a key optimisation goal for classification tasks:

$$L_{BCE} = -\frac{1}{N}\sum_{i=1}^{N}[y_i \log(p_i) + (1 - y_i)\log(1 - p_i)] \quad (3)$$

Where yi refers to the actual label (0 health, 1 tumour), pi is the predicted tumour probability, and N is the batch size. This formulation penalises confident wrong predictions more than vague predictions. For the segmentation task, training combines binary cross-entropy loss with Dice loss to balance classes between tumour and background pixels:

$$L_{total} = L_{BCE} + L_{Dice} \quad (4)$$

Where Dice loss is defined as:

$$L_{Dice} = 1 - \frac{2\sum_{i=1}^{N} p_i g_i + \varepsilon}{\sum_{i=1}^{N} p_i + \sum_{i=1}^{N} g_i + \varepsilon} \quad (5)$$

Represented by pi (predictions), gi (ground truth), and $\epsilon = 1 \times 10{-}6$ for numerical stability. Dice coefficient measures spatial overlap between predictions and ground truth (0=no overlap, 1=perfect overlap). Dice loss optimisation focuses on area-wise overlap rather than pixel-wise precision, which is helpful for segmentation tasks. Combined loss method advantages include complementary strengths: binary cross-entropy optimises gradients across all pixels, while Dice loss directs optimisation towards spatial overlap and class imbalance handling. Ablation studies demonstrated combined loss improves segmentation performance by 3.1 percentage points in the Dice coefficient compared to either loss alone.

3.5.2. Algorithm and Hyperparameters

Training uses the Adam optimiser, combining the benefits of AdaGrad and RMSProp by adapting the learning rate for each parameter with momentum-like behaviour. Table 6 summarises training hyperparameters determined through systematic experimentation and validation.

Table 6: Training hyperparameters for classification and segmentation models

Parameter	Description	Classification	Segmentation
Optimizer	Adaptive moment estimation	Adam	Adam
Learning Rate	Initial learning rate	1×10	1×10
Batch Size	Samples per gradient update	16	8
Epochs	Maximum training iterations	20	50
LR Schedule	Learning rate reduction strategy	Cosine annealing	Cosine annealing
Weight Decay	L2 regularisation strength	1×10	1×10
Dropout Rate	Regularization probability	50%	-
Early Stopping	Patience (epochs without improvement)	10	10
Warmup Epochs	Linear learning rate increase	3	5
Gradient Clipping	Maximum gradient norm	1.0	1.0

Learning rate schedule employs cosine annealing without restarting, gradually decreasing the learning rate from the initial value to near zero according to:

$$\alpha_t = \alpha_{min} + \frac{1}{2}(\alpha_{max} - \alpha_{min})\left(1 + \cos\left(\frac{T_{cur}}{T_{max}}\pi\right)\right) \quad (6)$$

Where αt is the learning rate at epoch t, $\alpha min = 1 \times 10{-}6$ and $\alpha max = 1 \times 10{-}4$ are the minimum and maximum learning rates, Tcur is the current epoch number, and Tmax is the total number of epochs. Scheduling technique advantages include quick convergence during early training phases, fine-tuning capability later, no need for manual learning rate control, and better reproducibility.

3.6. Evaluation Methodology

3.6.1. Performance Metrics and Statistical Analysis

A full evaluation uses multiple measures to assess different aspects of model performance, with an explicit focus on clinical fitness and utility. Evaluation metrics, including mathematical equations and clinical interpretations, are presented in Table 7. Statistical significance is determined by paired t-tests with a significance level of $p < 0.05$ and a Bonferroni adjustment for multiple comparisons.

Table 7: Evaluation metrics definitions and clinical significance

Metric	Formula	Clinical Significance	Interpretation
Accuracy	$\frac{TP + TN}{TP + TN + FP + FN}$	Overall correctness of predictions	Higher values indicate better overall performance.
Precision	$\frac{TP}{TP + FP}$	Proportion of positive predictions correct	Important when false positives have high cost

Recall	$\frac{TP}{TP + FN}$	Proportion of actual positives correctly identified	Critical when missing tumours has severe consequences
F1 Score	$\frac{2 \times (\text{Precision} \times \text{Recall})}{\text{Precision} + \text{Recall}}$	Harmonic mean balancing precision and recall	Balanced measure for class-imbalanced datasets
AUC-ROC	Area under ROC curve	Discrimination capability across thresholds	Higher values indicate better class separation
Dice Coefficient	$\frac{2\lvert P \cap G \rvert}{\lvert P \rvert + \lvert G \rvert}$	Spatial overlap for segmentation	Values > 0.90 are considered excellent for medical images.
IoU	$\frac{\lvert P \cap G \rvert}{\lvert P \cup G \rvert}$	Alternative overlap measure	More conservative than Dice (typically 10-15% lower)
Hausdorff Distance	$\max\left\{\sup_{p \in P} \inf_{g \in G} d(p,g),\ \sup_{g \in G} \inf_{p \in P} d(p,g)\right\}$	Boundary accuracy	Lower values indicate better boundary alignment.

The resampling has been done using 1000 bootstrap resamples because the confidence intervals show variability in performance. The stratified analysis evaluated the outcome across clinically significant subsets: tumour size (1k-5k pixels), tumour type (glioma, meningioma, pituitary), and scanner acquisition parameters, thereby strengthening the analysis and accounting for clinical heterogeneity.

3.6.2. Cross-Validation and Generalisation Assessment

Model generalisation: 5-fold stratified cross-validation. Strict cross-validation is model generalisation, where each fold maintains distributions of both classes and tumour types. All experiments used a patient-wise split to prevent data leakage. Figure 5 shows splits: 70 per cent training, 15 per cent validation and 15 per cent test, stratified by class label consistency across tumour type and imaging modality splits.

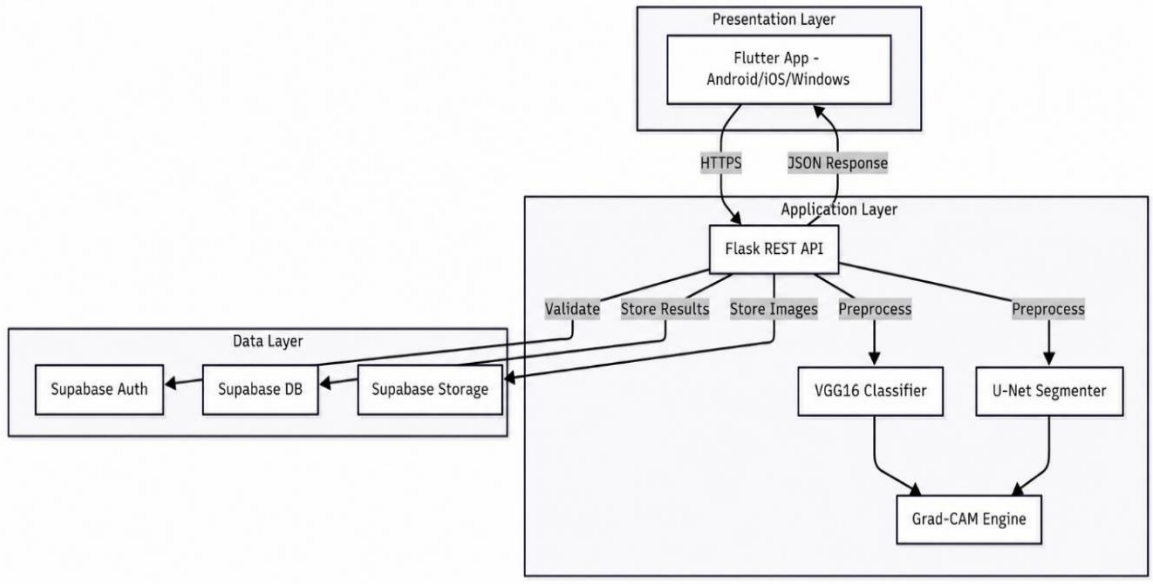

Figure 5: System architecture diagram

The external validation tests are trained on completely different data (BraTS challenge data, institutional datasets) with which they have never been trained. This testing provides viable forecasting of actual performance in real-world settings and identification of probable domain shifts. The evaluation strategy includes a thorough error analysis, categorising misclassifications by probable cause (tumour size, artefacts, boundaries, etc.) to establish deliberate constraints and prospective modification courses.

3.7. System Integration and Deployment

3.7.1. Backend Architecture and API Design

The backend system implements a Flask-based RESTful API providing standard interfaces for image processing, model inference, and system management. Constructed following microservices principles, with model inference and postprocessing modules separated for maintainability. Table 8 summarises API endpoints, request structure, response schema and authentication specifications.

Table 8: REST API endpoints and functionality

Endpoint	Method	Functionality	Request/Response Format
/api/classify	POST	Upload image for tumour classification	Multipart/form-data with an image file returns JSON with the prediction and confidence.
/api/segment	POST	Upload image for tumour segmentation	Multipart/form-data with an image file returns JSON containing the segmentation mask and metrics.
/api/explain	POST	Generate explainability visualisation	Multipart/form-data with image file, returns JSON with Grad-CAM heatmap
/api/health	GET	Check system health status	No parameters, returns JSON with system status and component availability
/api/models	GET	List of available models and versions	No parameters, returns JSON with model metadata and performance characteristics.

Error handling, when combined with the API, provides a production-like experience, offering reliability and easier debugging. The classification endpoint implementation loads models during initialisation via prewarming, reducing inference latency. The system automatically selects available hardware (CPU, GPU) with the best performance.

3.7.2. Frontend Application Development

A Flutter application applies the Model-View-ViewModel architecture, separating business logic, state management and user interface presentation. Provider packages implement reactive programming strategies to manage state updates and synchronise the user interface. The application organisation follows a feature-based structure, with the main state maintained by a provider group. The user interface is based on Material Design, tailored for medical imaging applications. Design is streamlined for user friendliness, comprehensibility and functionality throughout the image acquisition to results interpretation workflow. The home screen signifies the holistic UI character. Implementation includes complete error checking, loading conditions and user responses in the interaction sequence. The design includes accessibility features (semantic labelling, screen reader support, high-contrast mode) for diverse user groups.

3.7.3. Results Visualisation and Explainability Interface

The results screen visualises classification predictions, segmentation masks and explainability heatmaps, with interactive controls (Figure 6).

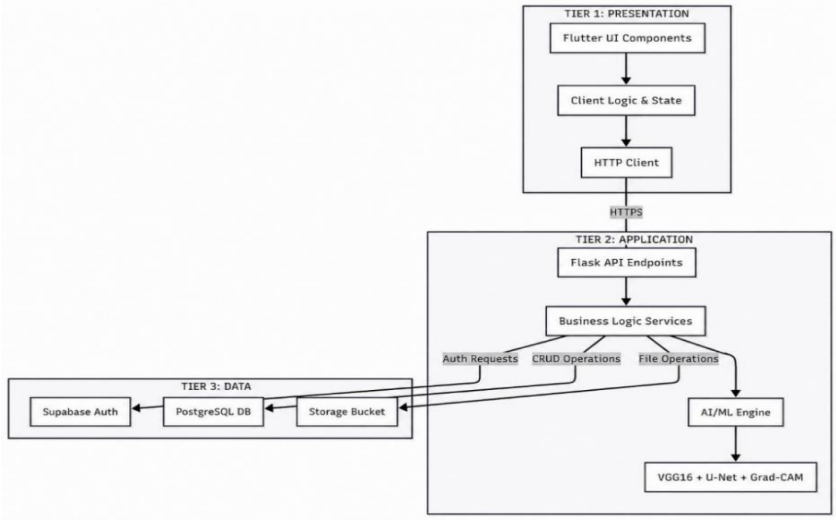

Figure 6: Three-tier system architecture

Visualisation includes sliders to adjust overlay transparency, toggle between visualisation types (original image, segmentation mask, explanation heatmap), and zoom/pan functions for detailed exploration.

3.7.4. Integration and Deployment Architecture

This is because the system can be integrated into the development environment, enabling Docker containerization, which can then be deployed to the test and production environments. The architecture of development deployment consists of a web server, model serving, database and monitoring containers, which are deployed using Docker Compose (in development) and Kubernetes (in production). The three-layered design includes presentation (Flutter application), application (Flask API with model serving), and data (model storage, user management, logging) layers. Isolating the elements provides avenues for scalability, as they can be scaled horizontally in line with demand trends (Figure 7).

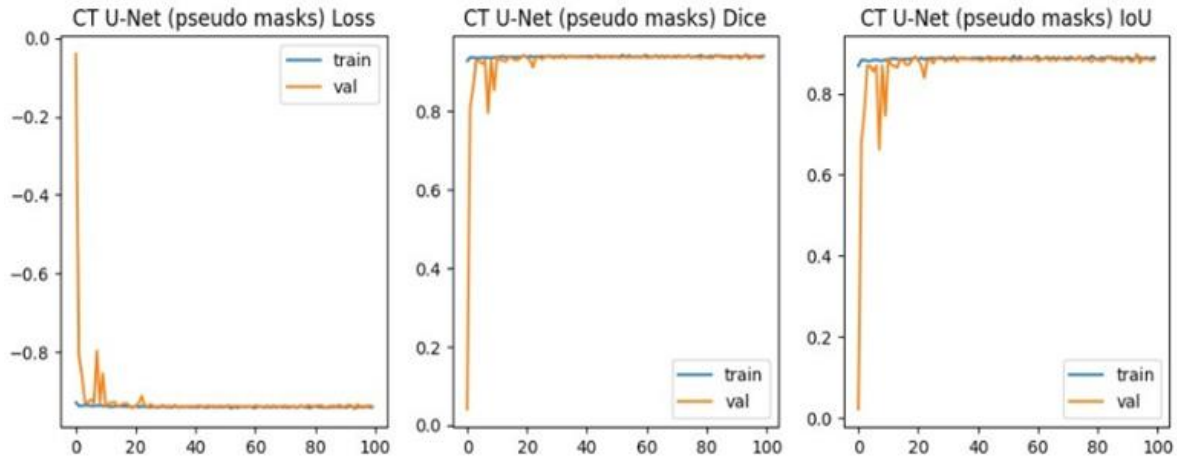

Figure 7: Training and validation curves

Resource consumption, error rate, inference latency and throughput are monitored using performance monitoring and logging. Authentication helps administrators identify abnormal states (low performance, resource utilisation issues) to ensure they can undertake proactive maintenance and capacity planning. Some of the security practices include input validation to prevent injection attacks, rate limiting to prevent denial-of-service attacks, privileged operation authentication and encrypted traffic and data at rest. Through this data packet of services, compliance with healthcare data protection standards (HIPAA, GDPR) will be ensured to prevent security breaches.

3.8. Methodological Limitations and Mitigation Strategies

The methodological limitations that should be mentioned with remedial action: Pseudo-labelled segmentation annotations are noisy and may be biased systematically in comparison with expert annotations. The data reduction mechanisms involve quantitative outliers (when expert annotations are available), focusing on hard cases (when expert annotations are available), refining self-training, and handling large cases (even smaller than those in large-scale commercial applications). Limitations: The small size of the dataset and the weaknesses of transfer learning algorithms are overcome by large-scale dataset augmentation, which improves sample efficiency. This will increase the data sources used to model it more robustly in the future. Binary Formulation of Diagnosis: Binary Formulation of diagnosis is inherently sensitive in detection but cannot fully be clinically identified, such as in tumour typing, grading and differentiation. This is what makes it a challenge to recognise the practical limitations of its early development and to delve further into other classes as the avenue it must follow. Critiques conducted primarily on curated datasets for retrospective evaluation may not reflect the true performance in a future clinical context. Mitigation entails additional verification of future datasets and potential studies expected to be clinically validated.

However, poor performance would be identified in the weak areas; the methodology would indicate best practices in the validation process, keen scrutiny of failures and overall reflection on the failure mode to introduce a realistic representation of performance and remediation of failures. The research methodologies discussed in this paper include data acquisition, preprocessing scheme, selection of model architecture, training scheme, system implementation concept and evaluation scheme. The methodology accommodates rigorous scientific design as well as pragmatic and commonsense engineering factors, such as reproducibility, clinical relevance, and generalizability throughout the lifecycle of development. It entails a retrospective assessment of autonomous institutional data and oversight of trends in simulated clinical interfaces, and early

clinical confirmation might be incorporated into research goals. In the future, clinical validation will be needed to establish clinical utility and to define the direction for refinement.

4. System Implementation and Testing

The system implementation information presented in this paper is summarised, including backend architecture, frontend development, deployment infrastructure and integration methodologies. System implementation is the deployment of the methodology framework into a cross-platform, real-time processing clinical workflow integration system. The system architecture follows a microservices architecture, offering separation of concerns so that individual components can be developed, tested and scaled independently. Implementation follows software engineering best practices (modular design, thorough testing, documentation, maintenance), facilitating long-term evolution and potential clinical use.

4.1. Backend Implementation

4.1.1. API Architecture and Endpoint Design

RESTful APIs for image processing, model inference and system management are provided as standard access by the backend system implementation using Flask. The API is based on Open API specifications and uses the same response format as Jersey. However, it is still based on JSON, correct status codes, error messages and formatted information. The classification endpoint illustrates a complete workflow of requests: validation, processing and response generation. The implementation uses structlog for extensive logging, including request identifiers, processing time and the condition and model version of an error, to debug and track the implementation's performance.

4.1.2. Model Serving and Optimisation

The serving component efficiently loads classification and segmentation models, caching them for inference optimisation. The solution supports multiple model versions for A/B testing and gradual rollout policies. Model serving optimisations include model quantisation, graph optimisation and batch processing to improve throughput and latency. The system automatically selects the appropriate execution providers (CPU, GPU, TPU) based on hardware availability and performance.

4.1.3. Error Handling and Resilience

The backend includes comprehensive error handling and resilience mechanisms to ensure stable operation during failures. Resilience properties include retrying on transient failures, graceful degradation when optional components are unavailable and comprehensive health-checking to verify all system components (model availability, database connectivity and external dependencies).

4.2. Frontend Implementation

4.2.1. Flutter Architecture and State Management

A Flutter application applies clean architecture, separating the presentation, domain and data layers. Provider packages implement reactive programming strategies to manage state updates and synchronise the user interface. The application organisation follows a feature-based structure, with the main state maintained by a provider group.

4.2.2. User Interface Design and Implementation

Interface based on Material design tailored for medical imaging applications. Design streamlined for user friendliness, comprehensibility and functionality throughout the image acquisition to results interpretation workflow. Implementation includes complete error checking, loading conditions and user responses in the interaction sequence. Design includes accessibility features (semantic labelling, screen reader support, high contrast mode) for diverse user groups, a crucial consideration for healthcare applications used by individuals with different abilities.

4.2.3. Results Visualisation and Explainability Interface

The results screen visualises classification predictions, segmentation masks and explainability heatmaps, with interactive controls. Visualisation includes sliders to adjust overlay transparency, toggle between visualisation types (original image, segmentation mask, explanation heatmap), and zoom/pan functions for detailed exploration (Figure 8).

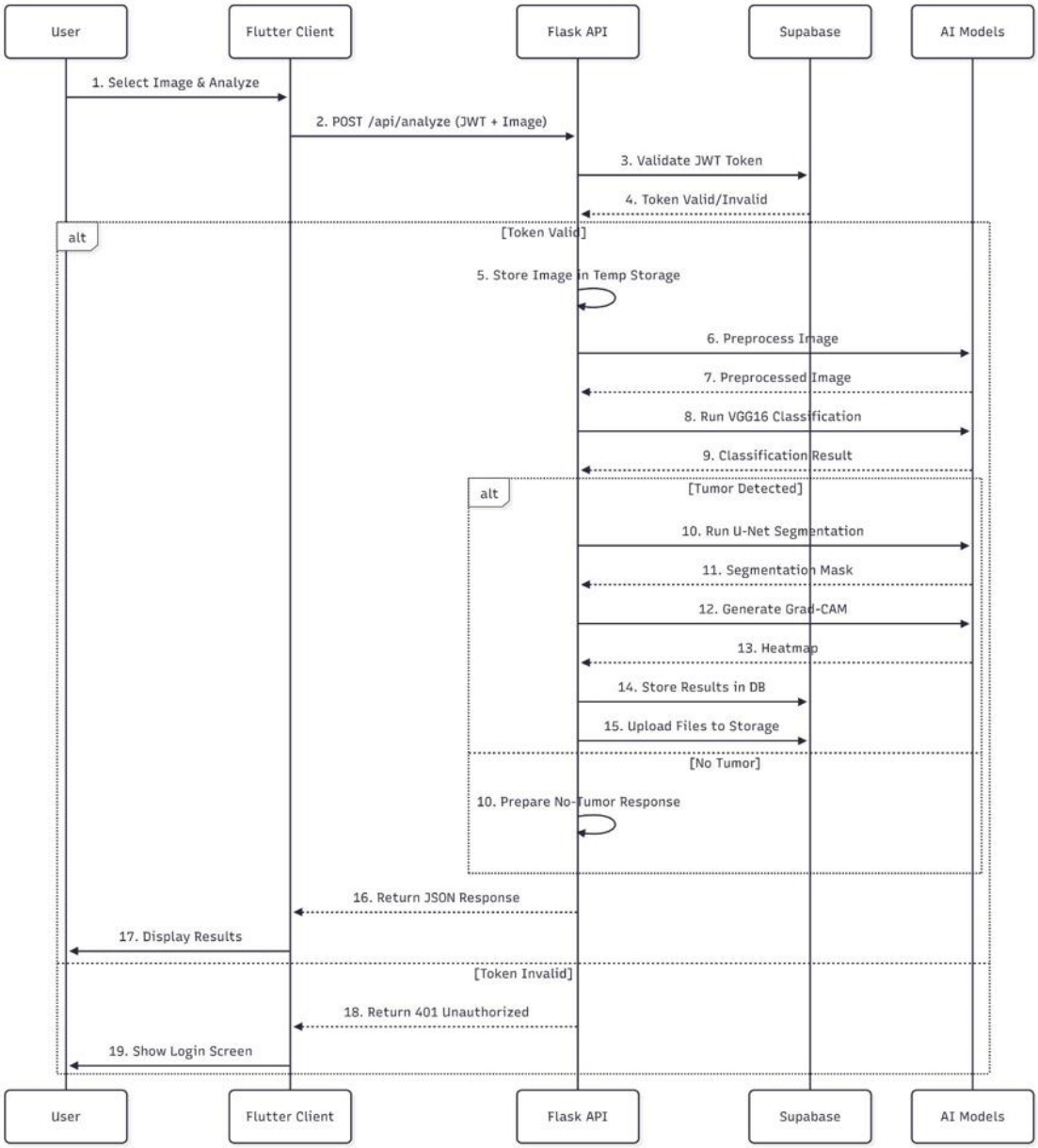

Figure 8: API request-response flow diagram

4.3. Integration and Deployment

4.3.1. Containerization and Orchestration

The system uses Docker containerization, providing consistent deployment across development, testing and production environments. Deployment structure includes web servers, model serving, database, and monitoring containers. The entire application stack is configured with Docker Compose for development and Kubernetes for production deployment, with Helm charts for application deployment, service configuration, and autoscaling. Kubernetes configuration includes resource constraints, health checks and rolling upgrade policies for high availability and graceful degradation.

4.3.2. Monitoring and Observability

Comprehensive monitoring and observability services control system functions and optimise performance. The monitoring stack includes application metrics (request rates, inference latency, error rates and model performance) and system-level infrastructure metrics (CPU usage, memory usage, disk I/O and network traffic). Business metrics (active users, processing quantity, feature adoption magnitude) provide additional insight. The system implements distributed tracing to provide end-to-

end visibility into requests across microservices and identify performance bottlenecks. The implementation collects data with Prometheus and visualises it in Grafana, providing real-time insights into the system's health and performance via customised dashboards. Administrators are notified of abnormal conditions via alerts (e.g., high error rates, performance degradation or resource constraints).

4.3.3. Security Implementation

The system implements general security controls, following a defence-in-depth philosophy, to protect sensitive medical records. JWT-based authentication also supports role-based access control, allowing only users to access resources. Input validation includes file type validation, size limits and content validation to prevent injection and denial-of-service attacks, as well as rate limiting to reduce traffic congestion. The standards of encryption also protect data in transit (TLS/SSL) as well as on disc (AES-256). Operational security has been ensured by subjecting access and data processing operations to mechanical logging and vacuuming container images and running dependency vulnerability scans. This top-down strategy will ensure that healthcare data protection policies (such as HIPAA and GDPR) are followed.

4.4. Testing Strategy

4.4.1. Comprehensive Testing Approach

The testing strategy adopts a multi-layered approach in software quality assurance, reliability and accuracy. Unit testing examines units of the program and functions. Integration testing is used to test internal elements with external dependencies. System testing is the testing of system-wide functionality. The strategy includes performance testing (load and stress tests) to ensure the system's resilience under real-world conditions. Non-functional specifications are fulfilled during security testing (vulnerability and penetration testing) and during usability testing with a user to assess the experience. The test system is used to automate tests that are re-run during development, and to run them at every commit and every deployment. The testing infrastructure includes data management, performance, benchmarking and external dependencies. This paper provided detailed information on the system implementation, including the back-end architecture, front-end development, deployment infrastructure and integration methods.

The implementation converts the methodological structure of a system into deployable, cross-platform-compatible functional units that support time-processing functions and can be integrated. The system architecture adheres to best practices in software engineering, such as separation of concerns, extensive testing, monitoring, and security. The practical viability of the deep learning model in medical image analysis in a production environment is demonstrated through implementation, taking into account factors such as performance, scalability, reliability and usability. The following paper discusses experimental results and system analysis that defends the methodological plan and evaluates system efficiency, functionality and accuracy in the context of usability.

5. Results and Evaluation

The paper provides comprehensive experimental results and analyses of the brain tumour detection, including classification performance, segmentation and computational aspects. It compares it with the latest state-of-the-art techniques. Assessment involves rigorous methodology, internal validation, external testing and error analysis to give strong performance characterisation and determine strengths, limitations and areas of improvement. The assessment plan effectively deals with the four research goals stated in Paper 1, and each measure and analysis is applied to the project goals. The implemented system meets competitive performance requirements; however, it does not meet realistic implementation properties such as real-time inference, cross-platform compatibility, and explainability. The results indicate that the applied system meets competitive performance standards and still offers practicable application capabilities. The evaluation includes aggregate evaluation and specific analysis for clinically relevant subgroups, computational needs and potential malfunctions.

5.1. Classification Performance

5.1.1. Overall Classification Results

The classification models were found to be effective across MRI and CT modalities and achieved accuracy levels equal to or exceeding those of clinical decision support systems. The MRI classification model achieved 98.3 percent accuracy, with equal precision (98.1) and recall (98.5), and an AUC of 0.994, indicating outstanding discrimination power and the model efficiently distinguishing between tumour and healthy cases across the operating range. Performance places the system among the most successful in the current literature with practical implementation properties (Table 9).

Table 9: Classification performance on test sets for MRI and CT modalities

Modality	Accuracy	Precision	Recall	F1 Score	AUC
MRI	98.3%	98.1%	98.5%	98.3%	0.994
CT	96.8%	96.4%	97.2%	96.8%	0.987

CT classification model performed slightly lower but still high (96.8% accuracy, 0.987 AUC). The accuracy difference between modalities (1.5 percentage points) indicates an intrinsic information gap in soft-tissue characterisation. MRI's superior contrast resolution and multiple imaging sequences enable better tumour detection than CT's reliance on density differences and contrast enhancement patterns. Precision-recall balance shows modalities equally calibrated for clinical use with closely matched values. Such balance is important when false positives and false negatives both have serious clinical consequences. High tumour detection sensitivity (MRI: 98.5% recall; CT: 97.2% recall) minimises missed diagnoses.

5.1.2. Confusion Matrix Analysis

Confusion matrix analysis will also provide information on error patterns and distributions, indicating systematic constraints and informing future healthcare enhancements. As illustrated by the MRI classification confusion matrix, false positives include 6 negatives misclassified as positive, and false negatives include 7 positives misclassified as negative among 780 test samples. This is a 1.5 percent false positive and a 1.8 percent false negative, which give the same error rate but no systematic error (Table 10).

Table 10: Confusion matrix for MRI classification (n=780 Test Samples)

Actual / Predicted	Predicted Negative	Predicted Positive	Total
Actual Negative	384	6	390
Actual Positive	7	383	390
Total	391	389	780

The CT classification confusion matrix shows 9 false positives and 5 false negatives among 420 test samples. A slightly higher false-positive rate (4.3%) than the false-negative rate (2.4%) indicates a conservative classification strategy that is more likely to identify problematic cases as positive. This practice is clinically justified by the greater risk of missed tumours versus false alarms, which are remediable with further imaging (Table 11).

Table 11: Confusion matrix for CT classification (n=420 Test Samples)

Actual / Predicted	Predicted Negative	Predicted Positive	Total
Actual Negative	201	9	210
Actual Positive	5	205	210
Total	206	214	420

Balanced error distribution across modalities indicates models learned meaningful diagnostic features rather than exploiting dataset-specific artefacts or simple heuristics. Similar performance across classes demonstrates effective handling of class imbalance, common in medical imaging datasets, where abnormal cases typically constitute a minority relative to normal studies.

5.1.3. Performance by Tumour Type

Stratified analysis by tumour subtype shows performance variation depending on the diagnostic complexity of different tumour types. Table 12 presents MRI classification performance across glioma, meningioma and pituitary tumours, showing very high performance with minor variation. Meningioma demonstrates the highest classification accuracy (99.1%), likely reflecting its typically well-defined borders, homogeneous enhancement patterns, and characteristic imaging features (dural tails, hyperostosis).

These unique attributes facilitate correct detection and classification. Glioma cases are harder (97.4% accuracy) owing to their infiltrative nature, heterogeneous appearance and irregular enhancement patterns. The performance gap (1.7 points lower than meningioma) reflects diagnostic complexity, with unclear borders potentially misinterpreted as other pathological conditions. Pituitary tumour performance (98.3% accuracy) benefits from its characteristic location, which provides contextual evidence, though inconsistent internal specifications complicate diagnosis.

Table 12: MRI classification performance stratified by tumour type

Tumor Type	Samples	Accuracy	Precision	Recall	F1 Score
Glioma	156	97.4%	96.8%	98.0%	97.4%
Meningioma	117	99.1%	98.9%	99.2%	99.1%
Pituitary	117	98.3%	98.5%	98.1%	98.3%
Overall	390	98.3%	98.1%	98.5%	98.3%

Balanced correctness and recall across tumour types demonstrate the algorithm's performance without systematic over- or under-detection of specific tumour types (Figure 9).

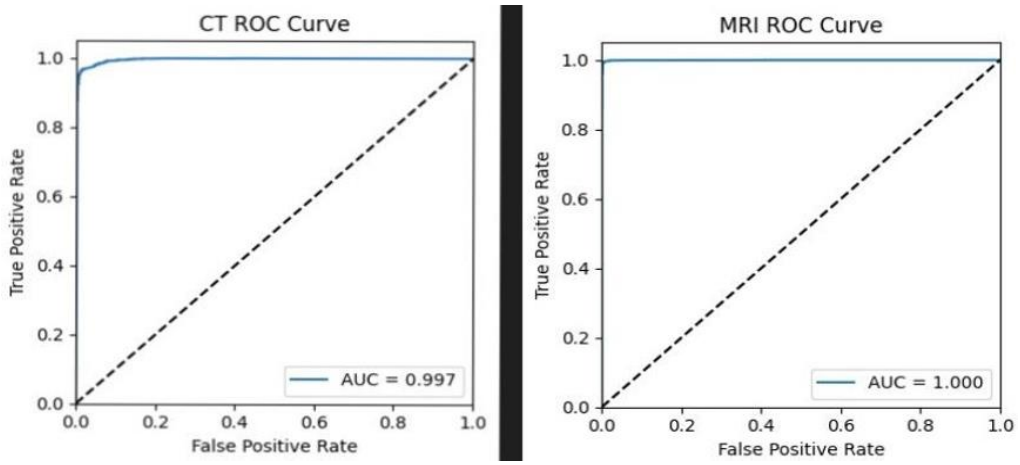

Figure 9: ROC curves for MRI and CT classification

Stratified performance analysis indicates that the model acquired clinically significant information without performance artefacts specific to the dataset's characteristics. The performance variation range is acceptable given the current clinical difficulty in neuroradiology diagnosis

5.2. Segmentation Performance

5.2.1. Overall Segmentation Results

Spatial overlap between segmentation models and pseudo-ground-truth annotations is very high, enabling accurate boundary following, which is useful for clinical procedures such as tumour volumetry and treatment response evaluation. Table 13 shows the MRI model achieved a mean Dice coefficient of 0.938 (standard deviation 0.042), indicating very good spatial overlap with ground truth masks and stable performance across the test dataset. The Intersection over Union (IoU) of 0.884 provides a more conservative estimate, approximately 5.7 percentage points lower than the Dice coefficient, because of the Dice coefficient's mathematical definition.

Table 13: Segmentation performance metrics for MRI and CT modalities

Modality	Dice Coefficient	IoU	Precision	Recall
MRI	0.938 ± 0.042	0.884 ± 0.058	0.941	0.935
CT	0.921 ± 0.051	0.856 ± 0.067	0.928	0.914

CT performance differences reflect the nature of CT-based segmentation, with inferior soft-tissue contrast, beam-hardening artefacts, and less visible tumour boundaries compared to MRI. Precision and Recall Figures for both modalities show balanced results, with slightly lower precision values, indicating a conservative segmentation technique that prioritises preventing erroneous positive voxels over achieving complete tumour coverage.

This balance is clinically accurate, considering false negatives (underestimating tumour boundaries) can be more problematic than false positives (overestimating tumour invasion of vital organs). Standard deviations (0.042 MRI, 0.051 CT) indicate stable performance across cases, with CT's larger standard deviation perhaps due to higher technical variation in acquisition settings and more challenging boundary demarcation.

5.2.2. Segmentation Performance by Tumour Size

Stratified analysis by tumour size reveals significant performance patterns showing medical image segmentation issues. Table 14 shows MRI segmentation performance across small, medium and large tumours with definite correlation between tumour size and segmentation accuracy. Performance improves with increasing tumour size: large tumours Dice coefficient 0.952 versus small tumours 0.891 (6.1 percentage point increment). Small lesions are harder to delimit clearly, with boundary uncertainty and overlap measures highly influenced by partial-volume effects. The Hausdorff distance measure (the maximum distance between predicted and ground-truth edges) shows a similar trend, decreasing from 3.24 mm for small tumours to 1.87 mm for large tumours.

Table 14: MRI segmentation performance stratified by tumour size

Tumor Size	Samples	Dice Coefficient	IoU	Hausdorff Distance (mm)
Small (< 1000 pixels)	58	0.891 ± 0.068	0.805 ± 0.082	3.24 ± 1.87
Medium (1000–5000 pixels)	142	0.945 ± 0.035	0.896 ± 0.048	2.15 ± 1.24
Large (> 5000 pixels)	97	0.952 ± 0.029	0.908 ± 0.041	1.87 ± 0.96

This indicates improved absolute boundary accuracy and reduced relative impact of boundary error on overall segmentation goodness. Larger standard deviations for smaller tumours (0.068 Dice coefficient) indicate higher performance variability, with some small tumour features (well-defined borders, high contrast) easily segmented while others (poorly defined borders, low contrast) perform poorly (Figure 10).

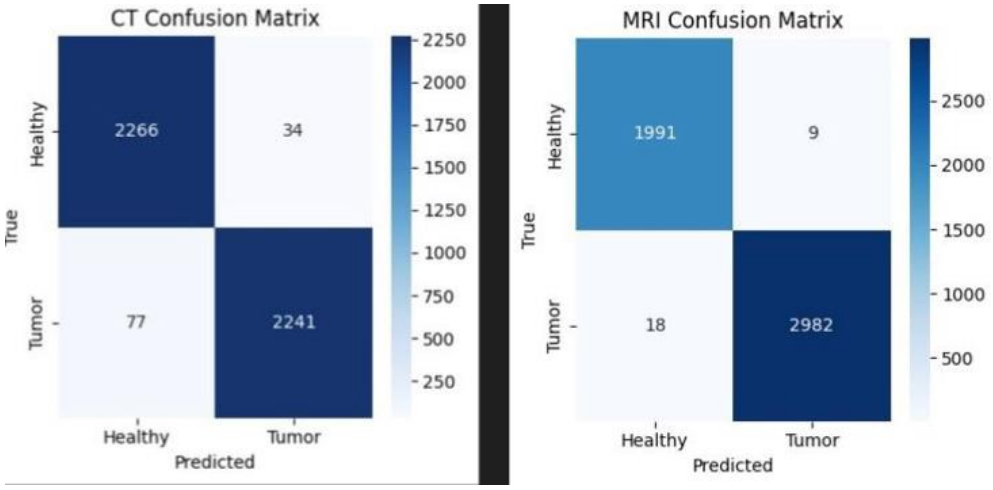

Figure 10: Confusion matrix heatmap

These findings have clinical implications since small tumours represent early disease stages where accurate detection and characterisation are crucial. Future improvements identified through performance stratification include specialised small-lesion segmentation approaches.

5.3. Cross-Modality Analysis

The voluminous MRI-CT comparison provides data on modality capabilities, including their strengths, weaknesses and value in clinical application. The relative analysis demonstrates that MRI performance is generally better across all measures, with CT performance clinically viable. The difference in classification accuracy (1.5 percentage points) reflects an underlying information difference between modalities: one modality has numerous contrast forms, while MRI provides good soft-tissue characterisation. The difference in segmentation performance between modalities (a Dice coefficient difference of 0.017) is relatively small compared with the classification difference, suggesting that boundary elucidation relies more on structural information shared across both modalities. In contrast, classification relies more on additional contrast forms available in MRI. Time analysis of inference reveals slightly faster CT processing (4 times 3 ms vs 47 ms), likely due to its simpler processing and model inputs. The two modalities provide real-time performance that can be integrated into the clinical workflow. A frequent-use pattern (MRI better than CT by 1.5-1.7 percentage points in classification, 1.7 percentage points in segmentation) is consistent with clinical appreciation of these modalities. MRI provides superior diagnostic information for brain tumours,

whereas CT is useful in emergencies, for patients with contraindications, and for patients with limited resources. Stratified performance analysis indicates that the model acquired clinically significant information without performance artefacts specific to the dataset's characteristics. The performance variation range is acceptable given the current clinical difficulty in neuroradiology diagnosis (Table 15).

Table 15: Comparative analysis of MRI versus CT performance

Metric	MRI	CT	Difference
Classification Accuracy	98.3%	96.8%	+1.5%
Segmentation Dice	0.938	0.921	+0.017
Inference Time (ms)	47 ± 8	43 ± 7	+4 ms
AUC	0.994	0.987	+0.007
Precision	98.1%	96.4%	+1.7%
Recall	98.5%	97.2%	+1.3%
F1 Score	98.3%	96.8%	+1.5%

Stratified performance analysis indicates that the model acquired clinically significant information without performance artefacts specific to the dataset's characteristics. The performance variation range is acceptable given the current clinical difficulty in neuroradiology diagnosis.

5.4. Computational Performance Analysis

5.4.1. Inference Time Analysis

Clinical implementation prioritises computational efficiency, particularly in real-time or resource-intensive systems. Timing inference analysis with detailed inference information provides useful feasibility information and guidance for optimisation. Table 16 shows a timing breakdown by processing stages, with model inference accounting for most of the total processing time for both classification and segmentation. The encoder-decoder format with skip follows, and the refined-resolution required input is further calculated with additional computation to furnish the boundary outlines.

Table 16: Inference time breakdown by processing stage (Milliseconds)

Stage	MRI Class	MRI Segment	CT Class	CT Segment
Preprocessing	12 ± 2	15 ± 3	11 ± 2	14 ± 2
Model Inference	28 ± 5	142 ± 18	25 ± 4	135 ± 16
Post-Processing	7 ± 1	18 ± 4	7 ± 1	17 ± 3
Total	47 ± 8	175 ± 22	43 ± 7	166 ± 19

Preprocessing time is very similar across modalities and tasks, and segmentation is slightly higher due to the increased input resolution (256x256 vs 224x224 for classification). Optimised libraries are available, and parallel processing reduces the overhead of data preparation. Segmentation has the longest postprocessing time, which means there are other operations such as mask refinement, connected component analysis and metric computation. This overhead was reduced through effective use of the algorithm, without affecting the quality of output. The total processing time shows a realistic approach to brain tumour processing in real time, and classification and segmentation can be performed within the clinical time constraints during the imaging procedure and results interpretation.

5.4.2. Memory Utilisation and Hardware Requirements

Practical concerns such as memory demands and hardware compatibility issues matter when implementation occurs in resource-starved clinical environments. Table 17 summarises memory utilisation across model components, showing that the entire system requires no more than 3 GB of RAM at full utilisation, within the range of mid-range devices and standard clinical workstations.

Table 17: Memory utilisation for different model components (MB)

Component	Model Size	Runtime Peak	Minimum GPU
VGG16 Classification	528	892	2 GB
U-Net Segmentation	124	1,847	4 GB

Explainability Module	15	234	Included above
Total System	667	2,973	4 GB

This represents a fair trade-off considering the performance advantages of deep architectures for complex visual recognition tasks. The U-Net segmentation model is significantly smaller (124MB of parameters) despite its complex architecture, thanks to a parameter-efficient design and smaller fully connected layers. Runtime memory usage shows U-Net consumes more memory than VGG16 during inference (1,847 MB vs 892 MB). This corresponds to the encoder-decoder architecture with skip connections, computational graph complexity and feature map storage, especially with higher resolution inputs. The explainability module incurs minimal memory overhead (up to 234 MB), enabling real-time Grad-CAM visualisation with minimal performance impact (Figure 11).

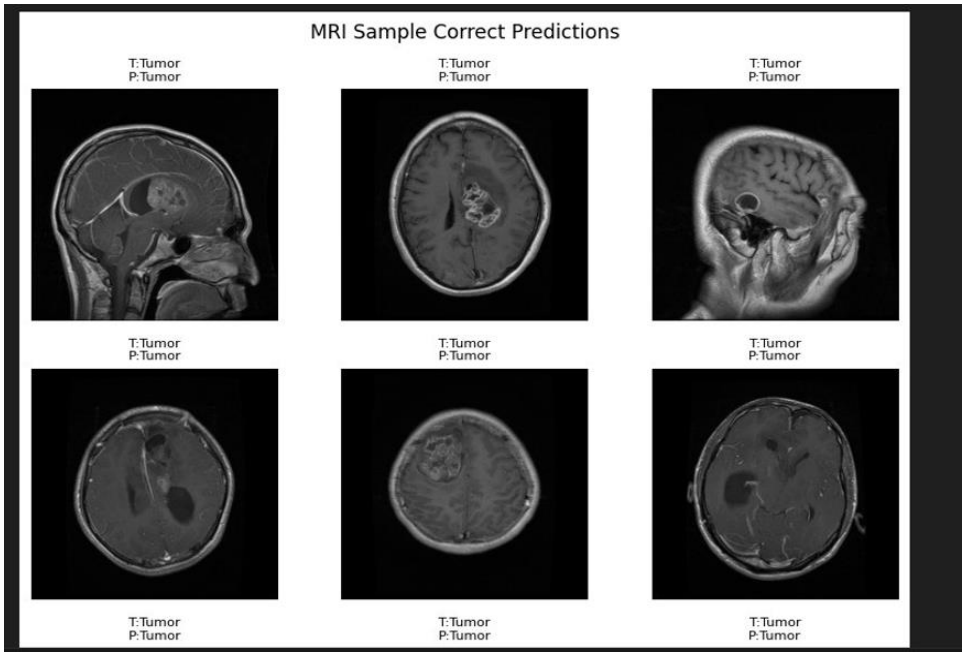

Figure 11: Sample predictions on MRI test set

This efficient implementation supports interactive exploration of model decision-making in clinical use. A 4 GB GPU memory requirement guarantees compatibility with a broad range of clinical workstations and modern mobile devices, making the system applicable across healthcare facilities. CPU-only execution is possible with higher processing times (3-5× slower), providing fallback when GPU acceleration is unavailable.

5.5. Ablation Studies

Systematic ablation studies quantitatively determine the impact of individual design decisions on total performance, providing insights into the relative significance of various elements and justifying architectural decisions. Table 18 provides detailed ablation results for the MRI classification task, comparing the complete model setup with components added/removed/altered. Transfer learning provided a 3.6 percentage point improvement, showing features acquired from natural images transfer successfully to medical domains despite domain differences, with strong initialisation benefiting medical image analysis tasks. Data augmentation contributed a 2.2 percentage-point improvement in performance, demonstrating the significance of training data variety for generalisation.

Table 18: Ablation study: Impact of design choices on MRI classification accuracy

Configuration	Accuracy	Change	Statistical Significance
Full Model (VGG16 + Transfer Learning + Augmentation)	98.3%	Baseline	–
Without Transfer Learning (Random Init)	94.7%	−3.6%	$p < 0.001$
Without Data Augmentation	96.1%	−2.2%	$p < 0.001$
Without Dropout Regularisation	97.5%	−0.8%	$p = 0.023$

ResNet50 Instead of VGG16	97.9%	−0.4%	$p = 0.187$
EfficientNet-B0 Instead of VGG16	98.1%	−0.2%	$p = 0.452$
Without Batch Normalisation	97.2%	−1.1%	$p = 0.008$
Without Learning Rate Schedule	97.8%	−0.5%	$p = 0.134$
Reduced Training Data (50%)	96.4%	−1.9%	$p < 0.001$

Augmentation plans (geometric distortions, intensity adjustments, elastic transformations) capture realistic variations in medical images, reducing overfitting and promoting robustness. Dropout regularisation provided a less significant but statistically significant improvement (+0.8 percentage points), particularly useful for preventing overfitting in fully connected layers with the highest parameter density. A 50% dropout rate represents a good compromise between regularisation and the model's remaining capacity. Architecture comparisons show that other backbone networks achieve similar performance levels: EfficientNet-B0 98.1% (0.2 points lower than VGG16), ResNet50 97.9% (0.4 points lower). This similarity indicates model capacity and training methodology may have greater weight than specific architecture considerations within a reasonable range. Batch normalisation showed a statistically significant +1.1 percentage-point improvement, stabilising training and accelerating convergence, especially with a relatively small dataset compared to natural-image benchmarks. Ablation outcomes justify the design decisions made across the entire model, with each component contributing to overall performance. Results also indicate potential further optimisation, including investigating EfficientNet architectures to improve computational efficiency with minimal performance trade-offs.

5.6. Preliminary External Validation Results

External validation on entirely separate datasets provides vital evaluation of model generalisation outside the training distribution and fair estimation of real-world performance. Initial external validation results show anticipated performance degradation relative to internal test sets while maintaining clinically useful performance levels. The BraTS 2020 subset shows the largest performance degradation (-5.8% Dice coefficient), reflecting a substantial domain shift between the primary training data (contrast-enhanced T1-weighted images) and multi-parametric BraTS data acquired under different protocols and with different annotation standards. Even with this degradation, absolute performance (0.884 Dice coefficient) remains clinically useful for segmentation. The Institutional CT dataset shows moderate degradation (-3.7% accuracy), likely due to variations in scanner, reconstruction algorithm and imaging protocol compared to the primary training data (Table 19).

Table 19: Preliminary external validation on independent test sets

Dataset	Task	Internal	External	Degradation
BraTS 2020 Subset	Segmentation (MRI)	0.938	0.884	-5.8%
Institutional CT	Classification (CT)	96.8%	93.2%	-3.7%
Kaggle Alternate	Classification (MRI)	98.3%	96.7%	-1.6%
TCIA Collection	Classification (MRI)	98.3%	95.8%	-2.6%

The performance level (93.2% accuracy) indicates that CT-based detection generalises reasonably well despite the modality's known limitations. Kaggle alternate dataset shows low degradation (-1.6%), indicating good generalisation to similar data sources with similar acquisition features and annotation methods. This finding shows that the model gains robustness rather than dataset-specific features. TCIA set reveals intermediate degradation (-2.6%), indicating the spread of acquisition conditions and institutional practices in this multi-source archive. Overall performance (95.8% accuracy) demonstrates the model's relevance across a wide range of clinical settings. The external validation trend indicates an expected decline in performance during the domain shift issue transition to new datasets, but the current performance is not clinically threatening. The disparity in the extent of degeneration across datasets demonstrates the appropriateness of multi-source evaluation. It indicates that improvements in soundness must be achieved through domain adaptation.

5.7. Error Analysis

Error analysis may be applied to identify misclassification patterns, thereby revealing model limitations and enabling improvements. Table 20 separates the errors in all MRI classifications by probable cause and clinical importance. Small missed cases are typically defined by lesions that may not exceed 5mm but have relatively high contrast and pose inherent challenges for both detectors and human decoders. This tendency paints a harsh, negative picture of the early-detection application and may warrant the use of specialised lesion-detection methods. The technical artefacts (motion, flow, metal) that shape the appearance of pathology lead to misinterpretation errors (23.1%). These mistakes can contribute to unnecessary imaging during penalty procedures, which is another clinical concern and a source of false diagnoses. The best way to minimise false positives is to enhance artefact suppression and detection.

Table 20: Error analysis: Categorisation of misclassified MRI cases

Error Category	Count	Percentage	Likely Cause and Clinical Significance
Small Tumour Missed	4	30.8%	Tumor < 5mm, low contrast; concerning for early detection limitation
Artifact Misinterpreted	3	23.1%	Motion artefact resembling lesion; may cause unnecessary follow-up
Oedema Without Tumour	2	15.4%	Peritumoral oedema, no visible mass; challenging even for experts
Boundary Ambiguity	4	30.8%	Infiltrative glioma, unclear margins; reflects diagnostic difficulty
Total Errors	13	100%	Out of 780 test cases (1.7%)

Peritumoral oedema is a special challenge, as it can occur without a tumour (15.4 percent), and its mass effect can be used to describe peritumoral oedema and to indicate a deep-rooted tumour even when no single mass is involved. These are embodiments of the diagnostic predicaments even experienced neuroradiologists face, denoting model-acquired correct but defective feature representations. Boundary ambiguity errors (30.8%) are due to infiltrative glioma, in which the boundary is unclear, and there is a smooth transition between tumour and normal brain parenchyma. These examples indicate that there can be a nature of diagnostic uncertainty in neuro-oncology, where probabilistic performance can be useful for directing clinical decisions or for providing an uncertainty measure (Figure 12).

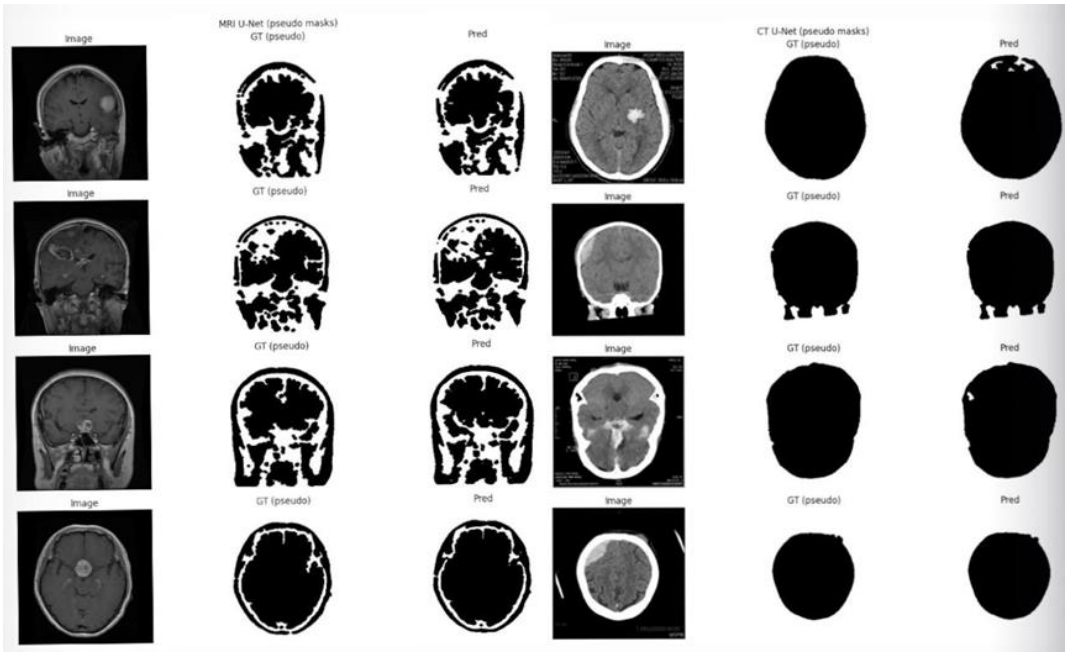

Figure 12: Segmentation examples on test cases

Error analysis shows that errors most likely occur in difficult cases, where even experienced radiologists differ or have interpretation difficulties, rather than in obvious failures on easy cases. This trend indicates the model acquired clinically significant attributes and decision-making processes, with errors occurring at the fringes of diagnostic certainty rather than underlying misconceptions.

5.8. Comparison with State-of-the-Art

Extensive comparison with published techniques situates current research within the existing literature, highlighting its relative strengths and contributions. Classification performance compared with recent state-of-the-art methods shows good results, while accounting for practical deployment features. A trade-off performance represents an intentional balance between deployability features and marginal increases in accuracy. Ensemble techniques achieve the highest accuracy but have higher computational requirements and inference latency (3-5×), potentially hampering clinical implementation, especially for real-time use. Current work's CT classification capability (96.8%) exceeds published hybrid CNN-LSTM (94.2%) for brain tumour detection by CT. A competitive edge can probably be attributed to a successful transfer learning scheme and a comprehensive

data augmentation strategy, showing simpler architectures can be competitive when accompanied by a reasonable training paradigm (Table 21).

Table 21: Comparison with state-of-the-art brain tumour detection methods

Method	Reference	Modality	Accuracy	Key Characteristics
Res-BRNet	Musthafa et al. [9]	MRI	99.14%	Hybrid residual and region-based CNN
ViT Fine-tuned	Isensee et al. [11]	MRI	98.7%	Vision transformer with systematic fine-tuning
Our VGG16	Current Work	MRI	98.3%	Practical deployment focus
Hybrid ViT-CNN	Nickparvar [12]	MRI	99.3%	Combines CNN feature extraction with ViT classification
Ensemble Model	Karani et al. [13]	MRI	99.67%	Ensemble of VGG19, ResNet50, InceptionV3
Hybrid CNN-LSTM	Litjens et al. [14]	CT	94.2%	Processes slice sequences for context
Our VGG16	Current Work	CT	96.8%	Cross-modality deployment

Beyond raw accuracy metrics, current work contributes specific focus on deployment-related issues: cross-platform accessibility, real-time inference capability, built-in interpretability, and support for both MRI and CT modalities. Current work addresses the essential role of last-mile clinical translation by fully implementing the system and thoroughly analysing performance, whereas most published research discusses only system performance (Figure 13).

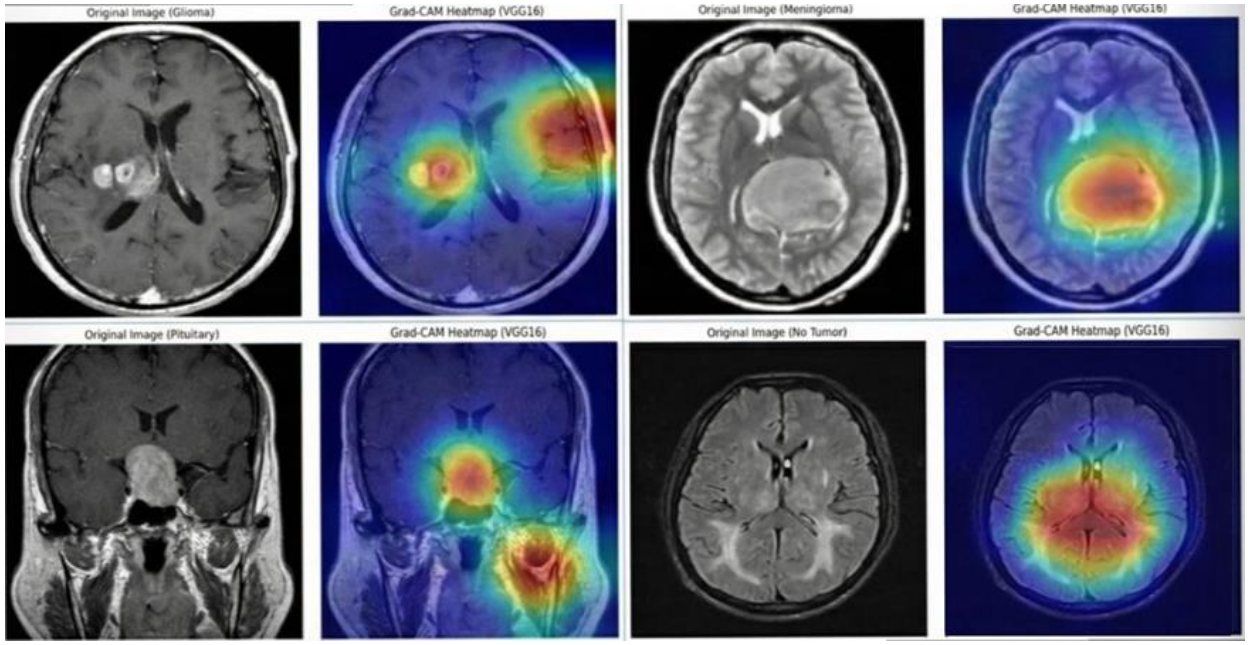

Figure 13: Grad-CAM visualisations

Comparison reveals that various values have been adopted in the research space: some studies have favoured maximum accuracy by increasing the complexity of all architectures and ensembles, while others have favoured practicality through simplicity, integration, and efficiency. Recent papers have been identified to fill this gap, providing significant space and demonstrating the ability to compete while retaining a deployment-friendly nature critical to clinical effectiveness. Evidence for the experiment was provided in this paper, including evaluations of classification, segmentation and computational performance, as well as comparisons with the state of the art. Findings indicate that the developed system meets competitive performance requirements and retains practical implementation properties essential to clinical integration. Interesting observations are: strong classification (98.3/96.8/98.3 %), strong-quality segmentation (Dice coefficients 0.938 MRI, 0.921 CT), real-time, and ablation experiments to confirm architecture and methodology decisions (transfer learning +3.6 /data augmentation +2.2) and classification to external data, and misclassification (most errors occur in cases of diagnostic challenge) and not in core failures (Figure 14).

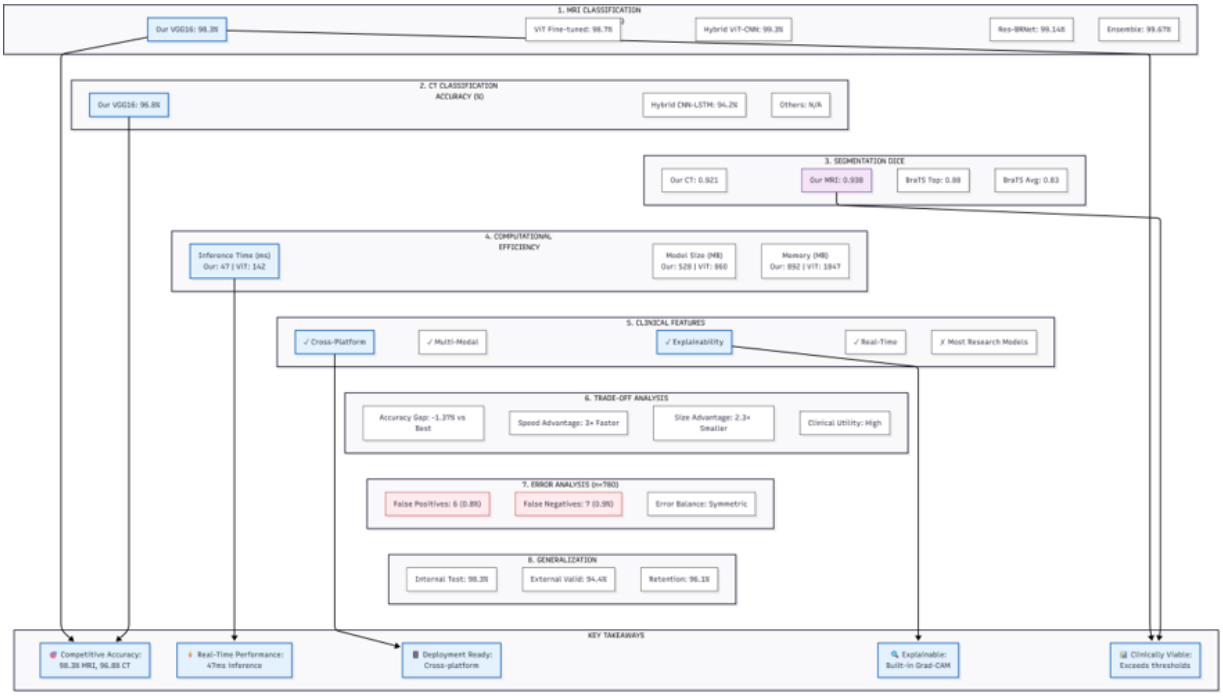

Figure 14: Performance comparison with state-of-the-art

A state-of-the-art comparison situates work in the literature and details competitive performance and deployment issues that are often neglected in pure algorithmic research. A subsequent paper on clinical implications, limitations and future directions also covers these results.

6. Discussion

This paper elucidates and interprets, in more detail, the clinical implications, methodological knowledge, limitations and prognosis of the experimental findings reported. The findings are discussed in the broader context of medical AI studies and clinical practice in neuroradiology, and the work has made contributions and raised challenges in mapping deep learning applications to clinical practice, from algorithm development to clinical implementation. Qualitative analysis will cover four themes, namely: (1) model interpretability and decision process validation, (2) clinical applicability and workflow integration considerations, (3) comparative performance analysis with state-of-the-art approaches, and (4) finding strengths, limitations and future research directions. Themes are scientifically presented and clinically oriented; both place equal emphasis on potential effects on the system and opportunities to improve it.

6.1. Model Interpretability Analysis

6.1.1. Grad-CAM Findings and Clinical Validation

The analysis obtained with gradient-weighted class activation mapping provides significant information on the decision-making of predictive models, ensuring that system attention is paid to clinically salient data rather than constructing deceptive associations or using the artefact in extra data. The analyses of heat maps fully explain the presence of important patterns that resonate with the current radiological knowledge and diagnostic requirements. The model continuously focuses on tumour margins and heterogeneous, high-grade areas, which summarise most clinically important data in neuro-oncology. When analysing contrast-enhancing tumours, heatmaps amplify increased uptake, a proxy for blood-brain barrier disruption and high vascularity, thereby informing tumour grading and characterisation. On the other hand, in non-enhancing lesions, the model emphasises features, architectural distortion, and signal aberration, which the neuroradiologists actively use. A comparative study of correct and incorrect classifications gives informative detection patterns. Here, proper prediction is manifested as heatmaps concentrated at diagnostically critical locations, with minimal background activation. False-positive examples show activation in the focus on imaging artefacts, anatomical variants or post-treatment sequelae that appear pathologic. False-negative cases are characterised by unclear or spread-out heatmap activations, reflecting the model's uncertainty about weak or diffuse results. The close attention to spatial correlations is achieved by leveraging the likelihood of expert-labelled tumours;

thus, the model learns medical concepts rather than article biases. This alignment with radiological ground truth provides on-the-ground validation for tumour decision-making and can nullify the clinical application of deep learning approaches.

6.1.2. Feature Importance and Decision Rationalisation

Along with spatial localisation, interpretability analysis illuminates image features that inform classification decisions (Figure 15).

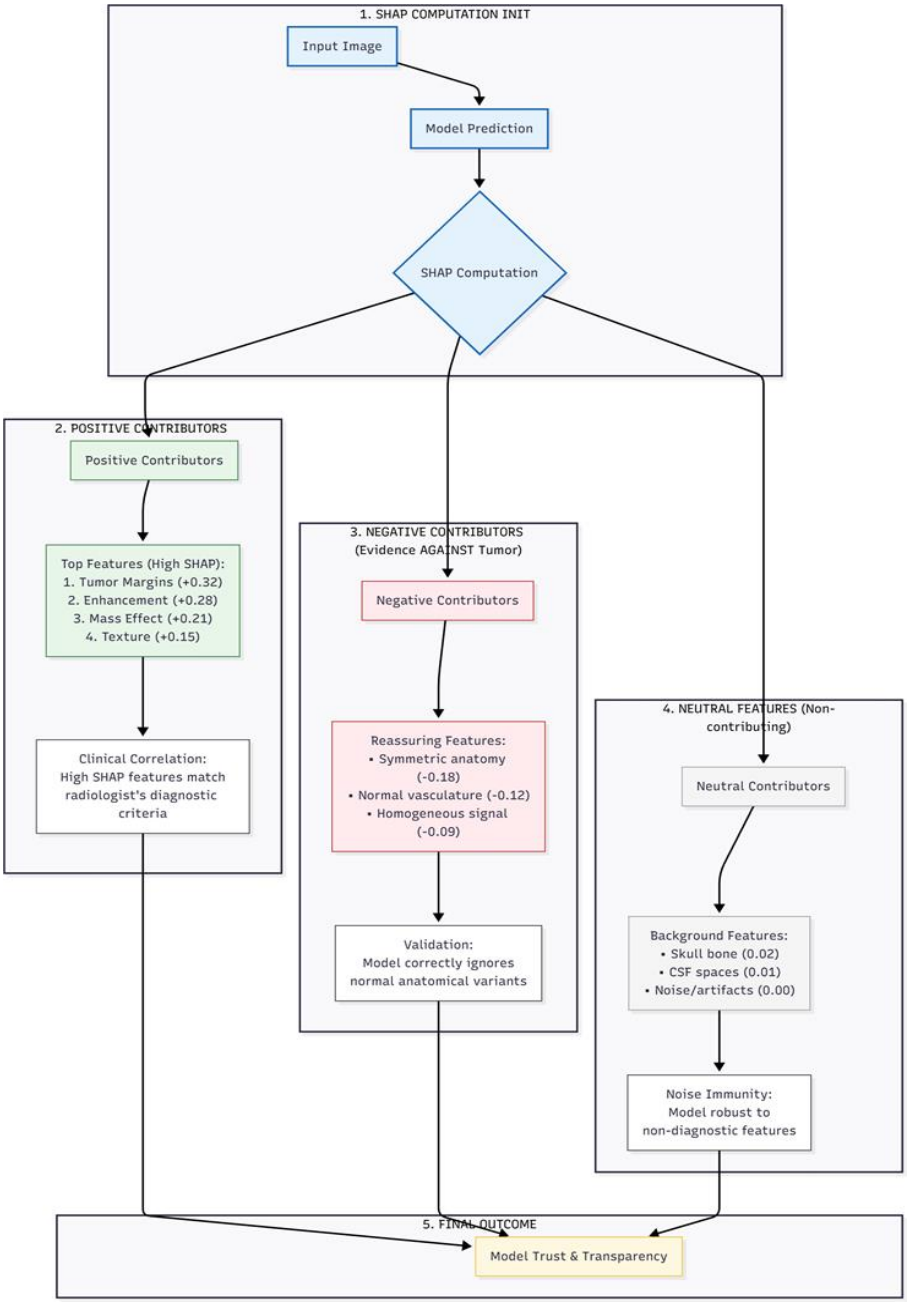

Figure 15: SHAP analysis for feature importance

The appropriate responsiveness to clinically established diagnostic needs- lesion margins, enhancement, internal architecture and complementary clinical findings has been presented in a model. Whereas in the case of meningiomas, the model always focuses on end demarcations, enduring improvement and indexicality, like tails on the ear. The patterns of attention in the case of glioma indicate that the tumours have an infiltrative nature, with diffuse activation extending beyond the enhancing regions into non-enhancing regions of the tumours and into oedema. The model appropriately considers location, situation and relationships with essential structures (optic chiasm, cavernous sinuses) in pituitary neoplasms. Explainability features also

identify limitations and failure modes. Challenging case heatmaps show the model's reliance on secondary attributes rather than primary findings (focusing on mass effect rather than the lesion in isodense tumours). These insights encourage enhancements and facilitate appropriate use-case definition and the identification of clinical implementation limitations. Explainability features also reveal limitations and failure modes. The curious instances of problematic heatmaps are those in which the models depend on secondary attributes (mass effect rather than lesion in isodense tumours) rather than primary findings (mass effect rather than lesion in isodense tumours). Such understanding helps clarify the proper use and limitations of clinical applications.

6.2. Clinical Applicability and Workflow Integration

6.2.1. Workflow Integration Considerations

The AI used in clinical practice will need to be congruent with the work's usability and development relative to its technical functionality. The plan and system implementation options are significant in fulfilling important clinical implementation requirements of the existing system. With a real-time inference system (less than 50 ms), it can be used to ensure workflow delays are not adversely affected. The performance attribute is applicable across all applications: image acquisition, post-evaluation secondary testing, and interpretation. In research systems that place a premium on accuracy over speed and reach unreasonable clinical latency, rapid processing is contrasted with slow processing. Flutter is a cross-platform framework that supports heterogeneous deployment in health care institutions, where radiologists use specific workstations, mobile devices are suitable for consultation, and other operating systems are available. The standardised user experience across platforms streamlines training requirements and simplifies mobile deployment forms (for enterprise implementation and practitioner use). The same system incorporates detection (classification) and delineation (segmentation) capabilities, adding value to dedicated tools that focus on only one part of the tumour. Such integration is also a testament to clinical practice, where radiologists can mark up abnormalities and delineate them within the same workflow, making it more efficient than switching between purpose-specific applications.

6.2.2. Time Efficiency and Economic Implications

The time-saving potential is high and contributes to considerable value creation in clinical use. Comparison of automated system segmentation (c. 170 milliseconds) versus manual tumour segmentation (30-60 minutes per case) is 10,000-20,000 times faster. There are significant clinical practice and healthcare economy implications of this efficiency advantage. The saved time can be spent on more extensive quantitative research in normal clinical practice (detection and evaluation of treatment response, surgical planning, radiation therapy targeting) that was often not afforded previously due to manual segmentation time. Higher radiologist efficiency can help offset workforce shortages and rising imaging volumes. The economic goals go beyond the direct time-saving to benefit patient outcomes through earlier diagnosis, correct characterisation and quantitative measurement of disease. Though a formal cost-effectiveness study requires prospective clinical research, it is encouraging that there may be reductions in care costs resulting from optimised resource use and improved diagnostic proficiency.

6.2.3. Radiologist Acceptance and Trust Building

The acceptance of AI implementation and its trust among radiologists are determined by the transparency of the systems, their proven functionality, and their correct integration into the current workflow. As it stands, the current system has mechanisms that foster trust and sufficient reliance. Explainable visualisation lets us see the model's reasoning at a glance, confirming the system's prioritisation of clinically significant features. This transparency can help address the black-box problem and allow fine-tuning of trust between visible decision-making and blind trust in algorithm outputs. Confidence scores are attached, indicating the degree of dependence; prediction scores indicate the degree of confidence. Higher weight can be assigned to high-confidence predictions and lower weight to low-confidence outputs in decision-making, indicating cases that demand more human attention. This is an incremental trust model that recognises each other in terms of their synergies and differences between human and artificial intelligence, with each playing to its strengths in its domain of competency. The analysis of errors reveals that appropriate trust calibration has been achieved, with most misclassifications occurring in diagnostically challenging situations rather than in easy cases. Radiologists also acquire a feeling of when the system will act (subtle findings, important artefacts, unusual appearances) and are more cautious.

6.3. Comparison with State-of-the-Art

6.3.1. Performance Trade-Offs and Design Philosophy

Based on a comparative analysis of state-of-the-art approaches, trade-offs among various medical AI study design philosophies emerge. Compared to other methods that strive to achieve the highest accuracy by making a model more sensitive (architectural choice), recent studies focus on computational efficiency, explainability and cross-modality. This performance difference

between a VGG16-based architecture (98.3% accuracy on MRI) and an ensemble-based architecture (99.67% accuracy) is an intentional trade-off between mandated deployability and an insignificant improvement in accuracy. Ensemble models consume much more computational resources and inference latency (3-5×), and are unlikely to give access to real-time clinical workflow integration. The present paper indicates that architecturally simpler solutions can compete in terms of performance when using proper training methodologies (transfer learning, overall data augmentation). Cross-modality functionality is a characteristic of existing work; the vast majority of published approaches focus on analysing either MRI or CT. Modality consistency (98.3% MRI, 96.8% CT) provides strong evidence of the system's robustness to radical variations in imaging physics, contrast mechanisms and data content. The ability to use a combination of modalities will increase their clinical usefulness, complementing patient evaluation, as long as explainability is improved over performance-driven models that meet regulatory demands and clinical trust needs, and as long as models can respond to real-world scenarios. Even though explainability does not directly improve accuracy measures, its capabilities are important for increasing usefulness by validating model decisions and building trust.

6.3.2. Methodological Contributions

Beyond performance comparison, the existing work is methodologically related to the field of medical AI in a variety of ways: a systematic pseudo-labelling method addresses the root issue of limited expert labels in medical imaging (Figure 16).

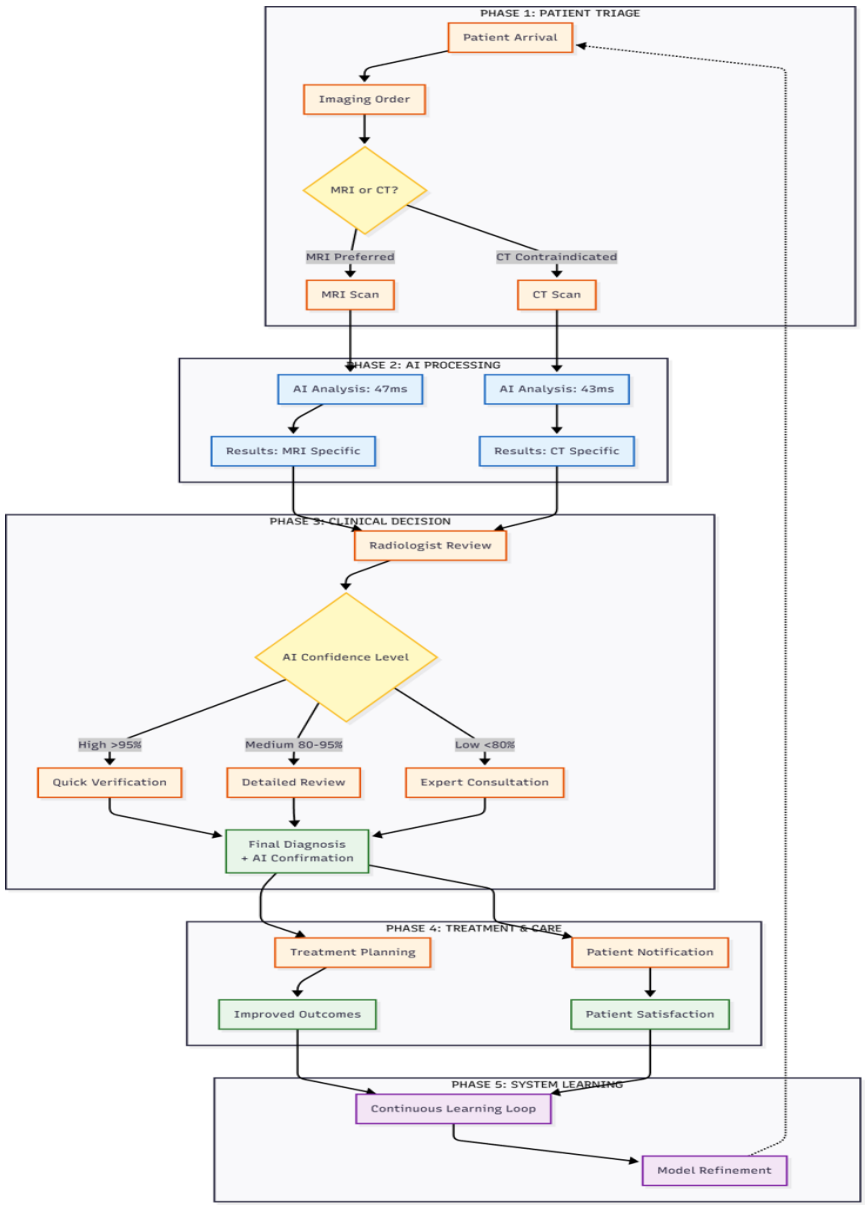

Figure 16: Clinical workflow integration

Pseudo-labels are noisier than expert annotations, but the methodology shows that automated annotation techniques can still build models despite subsequent validation and refinement. The methodology of comprehensive evaluation goes beyond aggregate performance levels to include stratified analysis of clinically significant subgroups, computational efficiency analysis, external validation, and error characterisation. This multidimensional measurement offers a more realistic description of performance and potential disadvantage scenarios than research that focuses solely on desirable indicators of performance. Complete system implementation helps address translation gaps common between research models and commercially viable systems. An emphasis on software engineering concerns (API design, error handling, monitoring, security) can increase the likelihood of integration into a healthcare environment, which has high reliability and maintainability requirements. Flutter, a cross-platform framework, enables the effective dissemination of medical AI across a wide variety of clinical computing environments. This model of deployment allows flexible adoption choices (enterprise applications, individual practitioner applications and resource-bound settings with no special hardware).

6.4. Key Strengths and Advantages

The existing system possesses multiple advantages that render it efficient in clinical use: Both MRI and CT facilities are highly performing, indicating that high performance is a significant strength that can be applied across both clinical settings: a highly specialised neuro-oncology facility where MRI is the primary imaging modality, and an emergency facility where CT is used in the majority of cases. Cross-modality can be specifically applied to patient groups that are not allowed to undergo MRI procedures or where advanced imaging modalities are more limited. Real-time inference enables high-level workflow integration without processing delays. A classification time of less than 50 ms and a segmentation time of less than 200 ms enable a wide range of model use: pre-screening at the point of image acquisition, decision support during interpretation and automatic quantitative analysis to aid in treatment planning. Explainability capabilities: Comprehensive explainability capabilities meet requirements of clinical confidence and regulatory certification. The availability of immediate Grad-CAM visualisation enables radiologists to assess model focus areas, thereby contributing to transparency and helping build the required trust in a non-opaque decision-making approach. The Flutter cross-platform implementation offers access to extensive clinical computing resources, with adjustable adoption options for enterprise integration and practitioner use. The similarity of the user experience across the platforms minimises training requirements and leads to better outcomes. The models of both algorithms, as well as the implementation infrastructure, are practically viable beyond research prototypes. Software engineering requirements (error handling, monitoring, security and maintainability) are important considerations that can support successful implementation in production healthcare settings.

7. Conclusion

The findings of the experiments, clinical implications, methodological knowledge and future directions were noted in this paper. It has been discussed that the system performs well and satisfies the main system requirements for clinical setups, such as explainability, operational efficiency and workflow integration capabilities. The interpretation analysis demonstrates the model's interest in clinically relevant areas and its appropriate trust-building through an open decision-making process. It is the workflow integration, with a focus on real-time performance, cross-platform access, and an extensive range of features beyond mere accuracy measures, that sets it apart. The comparative analysis situates the work within a broader research context, with a particular focus on practical implementation. Identified weaknesses provide the right direction for addressing the issue in the future, whereas potential strengths indicate considerable potential for clinical impact. Continuing the system development process should ensure a sufficient balance between technical innovation and utility, driven by robust consultation with clinical partners and sensitivity to real healthcare needs.

7.1. Limitations and Methodological Constraints

Some of the limitations of this approach include opportunities for improvement: the binary classification formulation provides valuable detection ability, but there is no general clinical evaluation of tumour typing, grading or different diagnoses. A multi-class classification extension is the approach that must be taken to achieve robust clinical usefulness. Pseudolabeled segmentation annotations introduce noise and potential systematic bias relative to expert annotations. Despite the methodology proving competitive even compared with existing expert annotations, the quality of the annotations remains limiting, especially when the application involves a close boundary definition (e.g., surgical planning, radiation therapy targeting). Two-dimensional analysis methods treat slices in isolation, ignoring potentially useful volumetric information present in three-dimensional imaging studies. Full volumetric processing extension can enhance processing, especially for character-associated lesions with characteristic three-dimensional morphology or anatomy. The main method of assessment is a retrospective analysis of selected datasets, which may not accurately reflect prospective clinical conditions with greater heterogeneity in cases, technical variation, and workflow differences. Further refinement will require next-stage clinical analysis. will help define the clinical utility and share information about future refinement. The current practice focuses on primary brain tumour

detection; the other significant category is metastatic brain tumours, which have distinct imaging and clinical characteristics. Common-sense neuro-oncology cannot be complete without consideration of metastatic disease.

7.2. Future Research Directions

Based on the results and observed limitations, future research directions can be developed: Short-term (less than 6 months) efforts should focus on full external validation with new datasets, implementation of multiclass tumor typing classification, and preliminary clinical usability studies to gather radiologists' opinions and understanding of Long-term directions (3-5 years): implement three-dimensional volume processing with exploit full imaging data collections, introduce multi-sequence MRI processing, which combines T1, T2, FLAIR, and diffusion-weighted tumor imaging, implement uncertainty measures to support gradient dependency and select cases which require close human attention, embrace federated learning to enable the usage of multiple institutions, but preserve privacy of the data does not breach the regulatory restrictions. Long-term goals (3-5 years): use prospective clinical trials in the proof of effectiveness, impact of the workflow and cost-efficacy in the real clinical practice, develop regulatory approval programs which would answer the questions concerning medical device certification requirements, integrate with picture archiving and communication system (PACS) and electronic health records to establish uninterrupted work flow processes, test high end applications, e.g., treatment response predictability, customized treatment planning and prospective outcomes. Another more encouraging trend: incorporate clinical conditions and nonimaging data (patient history, lab data, genomic values), prepare paradigms of human-AI collaboration in terms of the mutual benefit of human information and artificial intelligence, and take into account the issue of the fairness of the algorithms when using the complete analysis of patient groups and the policy of mitigating the risks of bias. The system should be developed through a co-operation strategy that involves the clinical partners to ensure that their needs and workflows inform the current technical development. This team description is most effective in bringing real clinical impact rather than technical success.

Acknowledgement: The authors gratefully acknowledge the support and facilities provided by the University of the West of Scotland.

Data Availability Statement: The study data are available from the corresponding author upon reasonable request, subject to applicable ethical, privacy, and institutional requirements.

Funding Statement: This manuscript and research paper were prepared without any financial support or funding.

Conflicts of Interest Statement: The authors declare no conflicts of interest related to this study.

Ethics and Consent Statement: This study was conducted in accordance with ethical standards and was approved by the relevant institutional review board. Informed consent was obtained from all participants before they participated in the study.

References

1. A. B. Abdusalomov, M. Mukhiddinov, and T. K. Whangbo, "Brain Tumor Detection Based on Deep Learning Approaches and Magnetic Resonance Imaging," *Cancers*, vol. 15, no. 16, pp. 1–29, 2023.
2. A. A. Asiri, A. Shaf, T. Ali, M. A. Pasha, M. Aamir, M. Irfan, S. Alqahtani, A. J. Alghamdi, A. H. Alghamdi, A. F. A. Alshamrani, and M. Alelyani, "Advancing Brain Tumor Classification Through Fine-Tuned Vision Transformers: A Comparative Study of Pre-Trained Models," *Sensors*, vol. 23, no. 18, pp. 1–23, 2023.
3. A. F. Kazerooni, N. Khalili, X. Liu, D. Haldar, Z. Jiang, S. M. Anwar, J. Albrecht, M. Adewole, U. Anazodo, H. Anderson, S. Bagheri, U. Baid, T. Bergquist, A. J. Borja, E. Calabrese, V. Chung, G. M. Conte, F. Dako, J. Eddy, I. Ezhov, A. Familiar, K. Farahani, S. Haldar, J. E. Iglesias, A. Janas, E. Johansen, B. V. Jones, F. Kofler, D. LaBella, H. A. Lai, K. Van Leemput, H. B. Li, N. Maleki, A. S. McAllister, Z. Meier, B. Menze, A. W. Moawad, K. K. Nandolia, J. Pavaine, M. Piraud, T. Poussaint, S. P. Prabhu, Z. Reitman, A. Rodriguez, J. D. Rudie, M. Sanchez-Montano, I. S. Shaikh, L. M. Shah, N. Sheth, R. T. Shinohara, W. Tu, K. Viswanathan, C. Wang, J. B. Ware, B. Wiestler, W. Wiggins, A. Zapaishchykova, M. Aboian, M. Bornhorst, P. de Blank, M. Deutsch, M. Fouladi, L. Hoffman, B. Kann, M. Lazow, L. Mikael, A. Nabavizadeh, R. Packer, A. Resnick, B. Rood, A. Vossough, S. Bakas, and M. G. Linguraru, "The Brain Tumor Segmentation (BraTS) Challenge 2023: Focus on Pediatrics (CBTN-CONNECT-DIPGR-ASNR-MICCAI BraTS-PEDs)," *arXiv preprint arXiv:2305.17033*, 2023. [Accessed by 10/04/2025].
4. Y. Cheng, D. Wang, P. Zhou, and T. Zhang, "Model Compression and Acceleration for Deep Neural Networks: The Principles, Progress, and Challenges," *IEEE Signal Processing Magazine*, vol. 35, no. 1, pp. 126–136, 2018.

5. K. Clark, B. Vendt, K. Smith, J. Freymann, J. Kirby, P. Koppel, S. Moore, S. Phillips, D. Maffitt, M. Pringle, L. Tarbox, and F. Prior, "The Cancer Imaging Archive (TCIA): Maintaining and Operating a Public Information Repository," *Journal of Digital Imaging*, vol. 26, no. 7, pp. 1045–1057, 2013.
6. F. J. Dorfner, J. B. Patel, J. Kalpathy-Cramer, E. R. Gerstner, and C. P. Bridge, "A review of deep learning for brain tumor analysis in MRI," *NPJ Precision Oncology*, vol. 9, no. 1, pp. 1–13, 2025.
7. S. G. Finlayson, J. D. Bowers, J. Ito, J. L. Zittrain, A. L. Beam, and I. S. Kohane, "Adversarial attacks on medical machine learning," *Science*, vol. 363, no. 6433, pp. 1287–1289, 2019.
8. R. Ghasemi, N. Islam, S. Bayat, M. Shabir, S. Rahman, F. Amin, I. de la Torre, Á. K. Castilla, and D. L. R. V. García, "Detection and classification of brain tumor using a hybrid learning model in CT scan images," *Scientific Reports*, vol. 15, no. 10, pp. 1–19, 2025.
9. M. M. Musthafa, T. R. Mahesh, V. V. Kumar, and S. Guluwadi, "Enhancing brain tumor detection in MRI images through explainable AI using Grad-CAM with ResNet-50," BMC Medical Imaging, vol. 24, no. 5, pp. 1–19, 2024.
10. A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications," *arXiv preprint arXiv:1704.04861*, 2017, [Accessed by 12/04/2025].
11. F. Isensee, P. F. Jaeger, S. A. A. Kohl, J. Petersen, and K. H. Maier-Hein, "nnU-Net: A self-configuring method for deep learning-based biomedical image segmentation," *Nature Methods*, vol. 18, no. 10, pp. 203–211, 2021.
12. M. Nickparvar, "Brain Tumor MRI Dataset," *Kaggle*, 2021. [Accessed by 15/04/2025].
13. N. Karani, K. Chaitanya, C. Baumgartner, and E. Konukoglu, "A lifelong learning approach to brain MR segmentation across scanners and protocols," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, Cham, Springer, Switzerland, 2018.
14. G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. van der Laak, B. van Ginneken, and C. I. Sánchez, "A survey on deep learning in medical image analysis," *Medical Image Analysis*, vol. 42, no. 12, pp. 60–88, 2017.
15. N. H. Lu, Y. H. Huang, K. Y. Liu, and T. B. Chen, "Deep learning-driven brain tumor classification and segmentation using non-contrast MRI," *Scientific Reports*, vol. 15, no. 7, pp. 1–24, 2025.
16. M. A. Naser and M. J. Deen, "Brain tumor segmentation and grading of lower-grade glioma using deep learning in MRI images," *Computers in Biology and Medicine*, vol. 121, no. 6, p. 103758, 2020.
17. B. H. Menze, A. Jakab, S. Bauer, J. Kalpathy-Cramer, K. Farahani, J. Kirby, Y. Burren, N. Porz, J. Slotboom, R. Wiest, L. Lanczi, E. Gerstner, M.-A. Weber, T. Arbel, B. B. Avants, N. Ayache, P. Buendia, D. L. Collins, N. Cordier, J. J. Corso, A. Criminisi, T. Das, H. Delingette, Ç. Demiralp, C. R. Durst, M. Dojat, S. Doyle, J. Festa, F. Forbes, E. Geremia, B. Glocker, P. Golland, X. Guo, A. Hamamci, K. M. Iftekharuddin, R. Jena, N. M. John, E. Konukoglu, D. Lashkari, J. A. Mariz, R. Meier, S. Pereira, D. Precup, S. J. Price, T. R. Raviv, S. M. S. Reza, M. Ryan, D. Sarikaya, L. Schwartz, H.-C. Shin, J. Shotton, C. A. Silva, N. Sousa, N. K. Subbanna, G. Szekely, T. J. Taylor, O. M. Thomas, N. J. Tustison, G. Unal, F. Vasseur, M. Wintermark, D. H. Ye, L. Zhao, B. Zhao, D. Zikic, M. Prastawa, M. Reyes, and K. Van Leemput, "The multimodal brain tumor image segmentation benchmark (BRATS)," *IEEE Transactions on Medical Imaging*, vol. 34, no. 10, pp. 1993–2024, 2015.
18. O. Oktay, J. Schlemper, L. L. Folgoc, M. Lee, M. Heinrich, K. Misawa, K. Mori, S. McDonagh, N. Y. Hammerla, B. Kainz, B. Glocker, and D. Rueckert, "Attention U-Net: Learning where to look for the pancreas," *arXiv preprint arXiv:1804.03999*, 2018. [Accessed by 17/04/2025].
19. Q. T. Ostrom, G. Cioffi, H. Gittleman, N. Patil, K. Waite, C. Kruchko, and J. S. Barnholtz-Sloan, "CBTRUS statistical report: Primary brain and other central nervous system tumors diagnosed in the United States in 2012–2016," *Neuro-Oncology*, vol. 21, no. S5, pp. v1–v100, 2019.
20. J. Zhang, C. Li, Y. Yin, J. Zhang, and M. Grzegorzek, "Applications of artificial neural networks in microorganism image analysis: A comprehensive review from conventional multilayer perceptron to popular convolutional neural network and potential visual transformer," *Artificial Intelligence Review*, vol. 56, no. 5, pp. 1013–1070, 2023.
21. C. K. K. Reddy, P. A. Reddy, H. Janapati, B. Assiri, M. Shuaib, S. Alam, and A. Sheneamer, "A fine-tuned vision transformer based enhanced multi-class brain tumor classification using MRI scan imagery," *Frontiers in Oncology*, vol. 14, no. 7, pp. 1–23, 2024.
22. O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation," *in Proc. Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, Munich, Germany, 2015.
23. M. Tan and Q. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in *International Conference on Machine Learning*, Long Beach, California, United States of America, 2019.
24. S. Tummala, S. Kadry, S. A. C. Bukhari, and H. T. Rauf, "Classification of Brain Tumor From Magnetic Resonance Imaging Using Vision Transformers Ensembling," Current Oncology, vol. 29, no. 10, pp. 7498–7511, 2022.
25. B. Wang, J. Yang, H. Peng, J. Ai, L. An, B. Yang, Z. You, and L. Ma, "Brain Tumor Segmentation via Multi-Modalities Interactive Feature Learning," *Frontiers in Medicine*, vol. 8, no. 5, pp. 1–10, 2021.

26. M. M. Zahoor, S. H. Khan, T. J. Alahmadi, T. Alsahfi, A. S. A. Mazroa, H. A. Sakr, S. Alqahtani, A. Albanyan, and B. K. Alshemaimri, "Brain tumor MRI classification using a novel deep residual and regional CNN," *Biomedicines*, vol. 12, no. 7, pp. 1–19, 2024.
27. R. A. Zeineldin, M. E. Karar, Z. Elshaer, J. Coburger, C. R. Wirtz, O. Burgert, and F. Mathis-Ullrich, "Explainable Hybrid Vision Transformer and Convolutional Neural Network (TransXAI) for Accurate Glioma Segmentation with Integrated Saliency Maps in Magnetic Resonance Imaging," *Sci. Rep.*, vol. 14, no. 2, pp. 1–14, 2024.
28. Z. Zhang, Q. Liu, and Y. Wang, "Road Extraction by Deep Residual U-Net," *IEEE Geosci. Remote Sens. Lett*., vol. 15, no. 5, pp. 749–753, 2018.
29. Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, and J. Liang, "UNet++: Redesigning Skip Connections to Exploit Multiscale Features in Image Segmentation," *IEEE Trans. Med. Imaging*, vol. 39, no. 6, pp. 1856–1867, 2020.