Brain Tumor Analysis System

AI-Powered Segmentation and Clinical Interpretation

Version 1.0

1 Introduction

The Brain Tumor Analysis System is an advanced AI-driven framework that integrates **YOLOv11-segmentation** for precise tumor detection in brain MRI scans with **Retrieval-Augmented Generation (RAG)** powered by **Mistral-8x7B** (via Groq API). The system generates a detailed PDF report containing a comprehensive clinical analysis and tailored treatment recommendations for detected brain tumors, leveraging a vectorized medical knowledge base grounded in peer-reviewed literature. By combining computer vision and natural language processing, the system automates tumor segmentation, provides clinically relevant insights, and produces professional reports to assist radiologists, researchers, and medical educators. The output is designed for preliminary analysis and requires validation by qualified medical professionals to ensure clinical accuracy. This document details the system's development methodology, research design, limitations, and supporting literature, providing a robust foundation for its application in medical research and clinical support.

2 Problem Statement

Brain tumors represent a critical medical challenge due to their diverse histopathological profiles, complex imaging characteristics, and significant impact on patient outcomes. Manual segmentation of tumors in MRI scans is labor-intensive, subject to inter-observer variability, and requires extensive expertise, often taking hours per scan. Additionally, translating segmentation results into actionable clinical insights demands integration of vast medical knowledge, which is time-consuming and prone to inconsistencies. Existing automated systems, such as those based on U-Net or earlier YOLO models, achieve high segmentation accuracy but lack comprehensive clinical interpretation or treatment recommendations. The need for an integrated system that combines accurate tumor segmentation with detailed, evidence-based clinical analysis and treatment planning is evident. This project addresses these challenges by developing a pipeline that leverages YOLOv11 for segmentation and RAG-based LLM for generating detailed PDF reports, enhancing diagnostic efficiency and supporting clinical decision-making in neuro-oncology.

3 Methodology

3.1 Research Design

The research adopts a quantitative, experimental approach to design and evaluate an AI-driven pipeline for brain tumor analysis. The system combines **YOLOv11-segmentation** for tumor detection with **RAG** and **Mistral-8x7B** for clinical interpretation, producing a detailed PDF report. The design involves iterative development, including data preprocessing, model training, validation, and performance evaluation against medical benchmarks. The pipeline is modular, allowing for future enhancements such as 3D MRI support or hospital system integration. The research evaluates segmentation accuracy using metrics like Intersection over Union (IoU) and clinical relevance of LLM outputs against established neuro-oncology guidelines.

3.2 Data Collection

The dataset comprises brain MRI scans from publicly available repositories, including the Brain Tumor Segmentation (BraTS) dataset (Menze et al., 2015) and The Cancer Imaging Archive (TCIA) (Clark et al., 2013). These datasets include T1-weighted, T2-weighted, and FLAIR MRI images in `.jpg` or `.png` format, accompanied by ground-truth annotations for tumor boundaries. Approximately 10,000 images are collected, covering various tumor types (e.g., gliomas, meningiomas, metastases). Medical literature, including peer-reviewed papers on tumor classification and treatment (Louis et al., 2021; Weller et al., 2021), is gathered to populate the RAG knowledge base, ensuring evidence-based clinical analysis.

3.3 Data Preprocessing

MRI images are preprocessed to ensure consistency and compatibility with YOLOv11. Images are resized to 640x640 pixels using bilinear interpolation to maintain quality. Intensity normalization is applied to standardize pixel values (0–1 range) across modalities. Ground-truth annotations are converted to polygon coordinates in YOLO format (normalized x, y coordinates). Medical literature is extracted using `pdfplumber`, segmented into 500-word chunks, and embedded using `sentence-transformers/all-MiniLM-L6-v2` for storage in ChromaDB. Data augmentation techniques, such as rotation and flipping, are applied to increase dataset diversity and improve model robustness.

3.4 Model Development

The system is developed as a modular pipeline with four core components:

- **Segmentation**: YOLOv11-segmentation (Ultralytics) detects and delineates tumors, outputting binary masks and polygon coordinates. OpenCV is used for post-processing (e.g., green mask and red outline visualization).
- **Knowledge Base**: ChromaDB stores vectorized medical literature for RAG, enabling retrieval of relevant context (e.g., tumor characteristics, treatment protocols).
- **LLM Interpretation**: Mistral-8x7B (via Groq API) combines segmentation data and retrieved context to generate a detailed clinical analysis, including tumor type, symptoms, and treatment recommendations.
- **PDF Generator**: ReportLab compiles segmented images and LLM summaries into a professional PDF report, formatted for clinical and research use.

3.5 Model Training

YOLOv11 is fine-tuned on the preprocessed MRI dataset using a pre-trained model (`yolov11.pt`). Training parameters include a batch size of 16, learning rate of 0.001, momentum of 0.9, and 100 epochs, optimized with the Adam optimizer. Training is conducted on a GPU-enabled system (e.g., NVIDIA RTX 3080) to handle computational demands. The RAG pipeline leverages pre-trained embeddings, requiring no additional training, but the knowledge base is iteratively updated with new literature to enhance LLM accuracy.

3.6 Model Evaluation

YOLOv11 performance is assessed using standard segmentation metrics:

- **Intersection over Union (IoU)**: Measures overlap between predicted and ground-truth masks (target: IoU > 0.85).
- **Mean Average Precision (mAP@0.5)**: Evaluates detection accuracy at IoU threshold of 0.5 (target: mAP > 0.90).

LLM outputs are evaluated for factual accuracy by comparing clinical analysis and treatment recommendations to peer-reviewed guidelines (e.g., Louis et al., 2021). A subset of 500 test images is used for quantitative evaluation, with qualitative review by medical experts for clinical relevance.

3.7 Validation

The system is validated using a hold-out test set (20% of the dataset, approximately 2,000 images) to assess generalizability across tumor types and imaging modalities. Five-fold cross-validation is performed to ensure robust segmentation performance. LLM-generated reports are validated against WHO tumor classifications and treatment protocols, ensuring alignment with clinical standards. Feedback from neuro-oncology experts is incorporated to refine report content.

3.8 Ethical Considerations

The system adheres to ethical guidelines to ensure responsible use:

- **Non-Diagnostic Purpose**: The system is designed for research and preliminary analysis, explicitly stating the need for radiologist validation to prevent misuse in clinical settings.
- **Data Privacy**: Only de-identified, publicly available datasets are used, complying with ethical standards (e.g., HIPAA-compliant TCIA data).
- **Transparency**: All reports include a disclaimer emphasizing AI-assisted analysis and the necessity of professional review.
- **Bias Mitigation**: Training data is balanced across tumor types, ages, and genders to minimize algorithmic bias, with regular audits to ensure fairness.
- **Accessibility**: The system is designed to be open-source, promoting equitable access for research communities.

4 Limitation and Scope

4.1 Limitations

- **Clinical Use**: The system is not intended for direct clinical diagnosis and requires validation by a qualified radiologist.
- **Data Dependency**: Segmentation accuracy depends on the quality and diversity of the training dataset, which may not cover all tumor subtypes.
- **2D Limitation**: The system processes 2D MRI slices, limiting its ability to analyze 3D tumor volumes.
- **LLM Accuracy**: The quality of clinical analysis relies on the RAG knowledge base, which may be incomplete for rare tumor types.
- **Computational Resources**: Training and inference require high-performance GPUs, which may limit accessibility in low-resource settings.

4.2 Scope of Research

The research focuses on developing an AI-driven system for 2D brain MRI tumor segmentation and clinical interpretation, targeting research and clinical support applications. The system integrates YOLOv11-segmentation and Mistral-8x7B with RAG to produce detailed PDF reports with clinical analysis and treatment recommendations. Future scope includes support for 3D MRI volumes, integration with hospital PACS systems, and real-time knowledge base updates to enhance clinical relevance.

5 Literature Review

Brain tumor segmentation has advanced significantly with deep learning models. U-Net (Ronneberger et al., 2015) introduced a convolutional architecture achieving IoU scores of 0.7–0.85 for medical imaging. YOLO variants, including YOLOv8 and YOLOv11, have improved real-time segmentation performance, with mAP scores exceeding 0.9 for object detection tasks (Jocher et al., 2023). RAG-based LLMs, such as those described by Lewis et al. (2020), enhance factual accuracy in medical text generation by retrieving relevant context, critical for clinical applications. The WHO classification of CNS tumors (Louis et al., 2021) provides a standardized framework for tumor grading, while Weller et al. (2021) outline treatment protocols for gliomas, emphasizing multimodal approaches (surgery, radiotherapy, chemotherapy). Recent studies (Bakas et al., 2017) highlight the BraTS dataset's role in benchmarking segmentation models. This project builds on these advancements by integrating YOLOv11 with RAG-based LLM to deliver comprehensive clinical analysis and treatment recommendations, addressing gaps in automated diagnostic support.

6 Timeline

Phase	Description
Month 1	Collect MRI datasets (BraTS, TCIA) and medical literature
Month 2	Preprocess images and annotations; build RAG knowledge base
Month 3–4	Train YOLOv11 model; validate segmentation performance
Month 5	Develop RAG pipeline and integrate Mistral-8x7B
Month 6	Implement ReportLab for PDF report generation
Month 7	Evaluate system performance (IoU, mAP, clinical accuracy)
Month 8	Conduct validation with expert feedback; finalize documentation

Table 1: Project Timeline

7 References

- Bakas, S., et al. (2017). Advancing The Cancer Genome Atlas glioma MRI collections with expert segmentation labels and radiomic features. *Scientific Data, 4*, 170117.
- Clark, K., et al. (2013). The Cancer Imaging Archive (TCIA): Maintaining and operating a public information repository. *Journal of Digital Imaging, 26*(6), 1045–1057.
- Jocher, G., et al. (2023). YOLO: A Decade of Object Detection and Segmentation. *arXiv preprint arXiv:2304.00501.*
- Lewis, P., et al. (2020). Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. *NeurIPS 2020.*
- Louis, D. N., et al. (2021). The 2021 WHO Classification of Tumors of the Central Nervous System. *Neuro-Oncology, 23*(8), 1231–1251.

- Ronneberger, O., et al. (2015). U-Net: Convolutional Networks for Biomedical Image Segmentation. *MICCAI 2015.*
- Weller, M., et al. (2021). EANO guidelines on the diagnosis and treatment of diffuse gliomas of adulthood. *Nature Reviews Clinical Oncology, 18*(3), 170–186.