Explainable and Transparent AI Models for Glioma Diagnosis from Brain MRI

Primary Supervisor: Prof. Ajitha Rajan, School of Informatics
Second Supervisor: Prof. Paul Brennan, Centre for Clinical Brain Sciences, Cancer Research UK, Edinburgh Centre
External Supervisor: Dr. Rishi Ramaesh, Consultant Radiologist, NHS Lothian

Abstract

Gliomas are the most common and fatal malignant brain tumours in adults. Diagnosis and treatment response evaluation in patients with gliomas are still highly dependent on neuroimaging. AI-based deep learning (DL) models have demonstrated transformative potential in medical diagnostics, including glioma diagnosis from brain MRI. The black-box nature of modern deep learning (DL) models makes it challenging to trust and understand the rationale behind their decisions, especially in high-stakes domains such as Glioma. Explainable artificial intelligence (XAI) techniques aim to increase the trustworthiness and transparency of a DL model's decision-making process by providing accessible interpretations. The project aims to develop a new transparent and explainable AI technique tailored to explain Glioma diagnosis from Brain MRI. Existing techniques have focussed on using techniques like GRAD CAM, saliency maps and perturbation-based techniques like LIME, SHAP and Occlusion sensitivity. However, these techniques fail to capture clinical concepts in their explanations. The technique designed in this project will aim to bridge this gap. The student will work closely with a clinician to understand what concepts to use in explanations. The accuracy of the designed AI model and the explanations will be compared against state-of-the-art techniques and evaluated by clinicians. Additionally, the model will be assessed for robustness and generalised to different MRI modalities and device settings in hospitals.

Introduction

Gliomas are the most common and fatal malignant brain tumors in adults. Diagnosis and treatment response evaluation in patients with gliomas are still highly dependent on neuroimaging. Despite active multimodal treatment, the prognosis remains poor. Advanced MRI functional imaging techniques can reflect the changes of lesions from different perspectives (cell proliferation, blood perfusion, brain metabolism, etc.). Artificial intelligence can aid the massive information from multimodal MRI images and help improve the accuracy of early diagnosis. However, the reliability and trustworthiness of AI models in this domain remains a challenge.

The black-box nature of modern deep learning (DL) models makes it challenging to trust and understand the rationale behind their decisions, especially in high-stakes domains such as medical diagnostics[1]. Explainable artificial intelligence (XAI) techniques aim to increase the trustworthiness and transparency of a DL model's decision-making process by providing accessible interpretations. There are two classes of XAI – post-hoc, applied to models post-training, and ante-hoc, designed to make models intrinsically explainable. Multiple works have shown a reduction in performance of popular XAI techniques when applied to specific medical domains instead of general computer vision tasks [2,3]. Post-hoc image XAI approaches like LIME [4] and GRADCAM [5] have been shown to routinely miss important clinical features in medical scans [6,7]. Textual approaches such as LLaVA-Med [8] have shown success in general diagnostics, but can fail to capture pathology-specific features which would have been

identified by a human expert during diagnosis [9]. While there are multiple possible reasons for this, including the poor quality of public medical datasets and their ground truth annotations [10], we hypothesize that a major contributor to this drop in performance is the lack of expert input when designing XAI techniques. The limitations of existing XAI techniques and challenges in explaining Brain Tumour was explored in our recent paper [11].

The proposed PhD project will aim at developing a transparent and explainable AI model for diagnosing Glioma from multimodal MRI while providing explanations that is accessible to clinicians. The technique designed in this project will aim to bridge this gap by designing a transparent AI model for diagnosing Glioblastoma from MRI with visual and textual explanations for the diagnosis. The explanations will help build trust in the AI model. The student will work closely with a clinician to understand what concepts to use in explanations. Causal approaches to explanation will also be explored.

After deigning the AI model and explanations, the PhD will explore and improve the robustness of the model using adversarial examples that alter the device settings and hospital parameters. The goal of this part of the research is to improve the generalisation of the AI model so it works with different device manufacturers and hospital settings.

Research Challenge

The research challenges that will be addressed in this project are as follows,

1. **SOTA models and XAI:** Evaluate current State-of-the-art (SOTA) MRI-based AI models for Glioma diagnosis and the quality of XAI techniques for explaining the diagnosis. Include prompt-based textual models like Llava-Med and counterfactuals that understand the relationship between the input features and output classification in the evaluation.
2. **Clinical concepts:** Consult clinicians to collect clinical concepts that can be used to explain Glioma diagnosis and predict tumour growth from MRI
3. **Model design:** Train a deep learning model with datasets that include the MRI modality to predict Glioma diagnosis and tumour while transparently explaining the diagnosis **with relevant clinical concepts.**
4. **Uncertainty quantification:** Develop uncertainty quantification metrics to associate model predictions with a degree of confidence, for clinical correlations. Explore Monte Carlo dropout and conformal analysis as a start
5. **Technical Robustness**: Assess robustness of the models with respect to incompleteness (eg. missing modalities in samples during testing), out-of-distribution test data from different device settings using techniques like knowledge distillation, weight space ensembling, adversarial testing.
6. **Evaluation against SOTA:** Assess the effectiveness of the designed AI model with clinical concepts against SOTA models and SOTA XAI techniques
7. **Evaluation with clinicians:** Conduct an evaluation study with clinicians to evaluate the explanations and accuracy of the Glioma diagnosis with clinical dataset

Data

An overview of MRI datasets and their modalities for Glioma is provided in the following paper – Abbad Andaloussi, M., Maser, R., Hertel, F., Lamoline, F. and Husch, A.D., 2024. Exploring Adult Glioma through MRI: A Review of Publicly Available Datasets to Guide Efficient Image Analysis. *arXiv e-prints*, pp.arXiv-2409.

The LUMIERE dataset is a good starting point -

Suter Y, Knecht U, Valenzuela W, Notter M, Hewer E, Schucht P, Wiest R, Reyes M. The LUMIERE dataset: Longitudinal Glioblastoma MRI with expert RANO evaluation. Sci Data. 2022 Dec 15;9(1):768.

Responsible AI/Ethical Considerations

The PhD student working on this project will primarily be working on public datasets, All members of the research team are required to follow Responsible AI and ethical practices. The PhD student is strongly encouraged to take the following course on Data Ethics, AI and Responsible Innovation from EFI – https://efi.ed.ac.uk/programmes/data-ethics-ai-and-responsible-innovation-free-short-online-course-from-the-university-of-edinburgh/

Data management plans and ethics processes as regulated by data processors like eDRiS will be followed for any clinical data that is used. The project, however, does not rely on the use of NHS data and any risks associated with that can be mitigated by using public data to train and test the model.

Expected Outcome & Impact

Transparent AI model for glioma diagnosis that provides explanations using clinical concepts to build trust and debug the model. Model output is accompanied by uncertainty quantification. Evaluation against SOTA models and with clinical experts.

References

1. Wadden, J. J. 2022. Defining the undefinable: the black box problem in healthcare artificial intelligence. Journal of Medical Ethics, 48(10): 764–768.
2. Pitroda, V.; Fouda, M. M.; and Fadlullah, Z. M. 2021. An Explainable AI Model for Interpretable Lung Disease Classification. In 2021 IEEE International Conference on Internet of Things and Intelligence Systems (IoTaIS), 98–103
3. Brunese, L.; Mercaldo, F.; Reginelli, A.; and Santone, A. 2020. Explainable Deep Learning for Pulmonary Disease and Coronavirus COVID-19 Detection from X-rays. Computer Methods and Programs in Biomedicine, 196: 105608
4. Ribeiro, M.; Singh, S.; and Guestrin, C. 2016. "Why Should I Trust You?": Explaining the predictions of any classifier. In arXiv, 1602.04938v3
5. Selvaraju, R. R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; and Batra, D. 2019. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. International Journal of Computer Vision, 128(2): 336–359.
6. de Vries, B. M.; Zwezerijnen, G. J. C.; Burchell, G. L.; van Velden, F. H. P.; Menke-van der Houven van Oordt, C. W.; and Boellaard, R. 2023b. Explainable artificial intelligence (XAI) in radiology and nuclear medicine: a literature review. Frontiers in Medicine, 10.
7. Rafferty, A.; Nenutil, R.; and Rajan, A. 2022. Explainable Artificial Intelligence for Breast Tumour Classification: Helpful or Harmful. In Interpretability of Machine Intelligence in Medical Image Computing, 104–123. Cham: Springer Nature Switzerland.
8. Li, C.; Wong, C.; Zhang, S.; Usuyama, N.; Liu, H.; Yang, J.; Naumann, T.; Poon, H.; and Gao, J. 2023. LLaVA- Med: Training a Large Language-and-Vision Assistant for Biomedicine in One Day. arXiv:2306.00890.

9. Lu, M. Y.; Chen, B.; Williamson, D. F. K.; Chen, R. J.; Ikamura, K.; Gerber, G.; Liang, I.; Le, L. P.; Ding, T.; Parwani, A. V.; and Mahmood, F. 2023. A Foundational Multi-modal Vision Language AI Assistant for Human Pathology. arXiv:2312.07814
10. Oakden-Rayner, L. 2017. Exploring the ChestXray14 dataset: problems. LaurenOakdenRayner.
11. Benedicte Legastelois, Amy Rafferty, Paul Brennan, Hana Chockler, Ajitha Rajan, Vaishak Belle. Challenges in Explaining Brain Tumor Detection. *TAS 2023: 21:1-21:8*