Initial Project Proposal Form (CN6000)

Unveiling Model Rationales: An Explainable AI Framework for Interpreting Black-Box Tumor Predictions for Convolutional Neural Networks

u2536809

Proposed Aim

The primary aim is to develop and evaluate an Explainable AI (XAI) framework that transforms opaque Deep Learning (DL) models into transparent and trustworthy decision-support systems for automated brain tumor diagnosis and screening.

Proposed Objectives

1. To establish and analyse the distinct mechanisms for localizing relevant images features and quantifying features contributions to the final CNN model prediction.
2. To investigate the fundamental trade-off between model accuracy and model transparency in high-stakes images, establishing a balanced rationale for the integration of the chosen XAI suite.
3. To quantitatively evaluate the precision of the visual XAI output.
4. To conduct quantitative analysis (Robustness testing) of the features contributions using XAI methodologies.
5. To implement and train the proposed CNN model using available MRI images Dataset and integrate it with the post-hoc XAI techniques to create unified Explainable hybrid model.
6. To utilize the interpretability insights provided by XAI to prune CNN model to maintain high accuracy on unseen data.
7. To reflect in the limitations of the implemented hybrid model and propose future research directions.

Draft of Rationale

The integration of Explainable AI into deep learning models for MRI-based brain tumor diagnosis is essential for clinical trust, technical reliability, and regulatory compliance. Models like CNN can predict brain tumors by scanning MRI images, but their "Black Box" nature limits safe medical use. XAI addresses this by making model decisions explainable and interpretable, showing which features play a vital role for tumor detection. AI systems ensure alignment with real tumor characteristics. This process not only builds clinician confidence and helps decision-making but also helps detect bias and improve model efficiency, enabling responsible and trustworthy deployment of AI in neuro-oncology.

Facilities Required

Dataset

- Brain tumor MRI Images that are publicly available.

Computational Requirement

- Local computer or university facilities.
- Cloud-based platforms such as Google Colab or Kaggle.

Software Requirements

- Deep learning Framework (Pytorch, Tenserflow, etc).
- Explainability libraries: SHAP, LIME, Grad-CAM implementations.

Supervisor

Maimoona Sharif