Explainable AI for Brain Tumor Detection

Manan Savla
Department of Information Technology
Sardar Patel Institute of Technology
Mumbai, India
manan.savla@spit.ac.in

Dhruvi Gopani
Department of Information Technology
Sardar Patel Institute of Technology
Mumbai, India
dhruvigopani@gmail.com

Mayuri Ghuge
Department of Information Technology
Sardar Patel Institute of Technology
Mumbai, India
mayuri.ghuge@spit.ac.in

Sheetal Chaudhari
Department of Information Technology
Sardar Patel Institute of Technology
Mumbai, India
sheetal_chaudhari@spit.ac.in

Chandrashekhar Ramesh Gajbhiye
Department of Applied Sciences & Humanities
Sardar Patel Institute of Technology
Mumbai, India
c_gajbhiye@spit.ac.in

***Abstract*—The accurate detection and classification of brain tumors are critical components of medical diagnostics. Recently, artificial intelligence (AI) techniques have gained traction in enhancing brain tumor identification. However, AI models often lack transparency, which is especially problematic in high-stakes fields like healthcare. This project proposes an Explainable AI (XAI) system specifically for brain tumor detection, offering clinicians clear, interpretable insights into model decisions. By employing sophisticated XAI methods, this system aims to improve trust, reliability, and clinical adoption of AI-driven tumor detection solutions.**

***Keywords*—Medical Image Processing, Explainable AI, Brain Tumor, AI in Heathcare, LIME**

I. INTRODUCTION

This project focuses on developing an Explainable AI (XAI) framework tailored for brain tumor detection, addressing the current limitations in interpretability within AI-based medical solutions. By combining advanced machine learning algorithms with XAI methodologies, this system is designed to produce accurate diagnostic predictions accompanied by clear explanations. This transparency is intended to foster greater trust and collaboration between AI systems and medical professionals. Key components of the project include data collection, preprocessing, development of a deep learning model, and the integration of XAI techniques. The system's efficacy will be evaluated on real-world brain tumor datasets, assessing accuracy, interpretability, and usability.

This project has the potential to make a substantial impact on the field of medical diagnostics, particularly in brain tumor detection. By offering clinicians interpretable insights into AI-generated predictions, this system can support informed decision-making, reduce dependency on subjective interpretations, and ultimately contribute to enhanced patient care and outcomes.

.

II. LITERATURE REVIEW

The proposed crime detection model in [1] combines CNN, RNN, and BERT. Initially, the dataset is benchmarked on audio using CNN and RNN. Decision-level fusion is employed on a multimodal sentiment analysis model, followed by text analysis using BERT. The combination of CNN and BERT achieves better performance than RNN. The authors plan to extend the model to include images and video.

[2] The paper introduces an XAI approach for enhancing the interpretability of deep learning models in classifying retinal OCT images. Techniques like class activation maps, attention-based mechanisms, and saliency maps are proposed to visualize the model's focus on critical image regions. Evaluation on a public OCT dataset demonstrates superior accuracy and interpretability compared to traditional deep learning models. The authors highlight the potential of their approach in improving diagnostic decisions for retinal diseases.

In [3], an XAI approach is presented for predicting drug sensitivity in cancer cell lines. The authors utilize a deep neural network to make predictions based on genomic features and employ Layer-wise Relevance Propagation (LRP) to generate feature importance scores for interpretability. Evaluation on a publicly available dataset shows superior performance compared to traditional machine learning models in terms of accuracy and interpretability. The authors suggest that their XAI approach can help identify crucial genomic features and contribute to personalized cancer therapies.

In their study, [4] propose a cascaded CNN approach for automatic liver and tumor segmentation from CT and MRI volumes. They utilize dilated convolutions and residual connections in a two-stage architecture to improve segmentation accuracy. The evaluation on publicly available datasets shows that their approach outperforms traditional segmentation techniques in terms of both accuracy and efficiency. The authors emphasize the practical applications of

their approach in computer-aided diagnosis and treatment planning for liver cancer.

In their study, [5] the authors propose an innovative approach that utilizes XAI techniques to understand the spatio-temporal dynamics of COVID-19 risk factors. They employ advanced machine learning algorithms to identify patterns and correlations among these factors. By using rule-based models and feature importance analysis, they extract meaningful insights and provide interpretable explanations. The study introduces the XAI4COVID model, which enables predictions of COVID-19 cases and deaths while highlighting the importance and impact of all factors.

In [6], a retinal disease classification model for OCT images is proposed. The authors aim to enhance interpretability and transparency by combining a deep neural network with XAI techniques. The study achieves high accuracy with CNN models and utilizes the LIME framework to explain misclassifications. Extensive evaluation, validation on diverse datasets, and integration of domain knowledge are highlighted for real-world applicability.

The study in [7] aims to enhance liver tumor detection in MRI images. The authors propose MIMFNet, a multimodal detection framework that combines local and global features from various scales and modalities for improved accuracy and localization. The evaluation confirms the effectiveness of the proposed framework. The authors suggest future directions, such as evaluating on larger datasets, incorporating additional clinical data, addressing computational efficiency challenges, and exploring interpretability and visualization techniques.

[8] The focus is on integrating Explainable AI (XAI) techniques in healthcare to enhance transparency and interpretability of AI models. Authors emphasize the importance of incorporating XAI methods into healthcare AI systems, providing interpretable explanations for predictions. Research gaps include standardized evaluation metrics, ethical guidelines, domain knowledge integration, and user-friendly interfaces. Future work involves developing communication tools, exploring ensemble methods, and evaluating XAI's impact on decision-making, patient care, and trust in AI systems.

The proposed crime detection model in [9] combines CNN, RNN, and BERT. Initially, the dataset is benchmarked on audio using CNN and RNN. Decision-level fusion is employed for multimodal sentiment analysis, followed by text analysis using BERT. The combination of CNN and BERT achieves better performance than RNN. Future plans include extending the model to incorporate images and video.

[10] The paper examines the progress of deep learning in medical image analysis and acknowledges the limited acceptance among medical professionals due to transparency issues. To tackle this challenge, researchers have developed explainable AI methods to enhance interpretability. The survey explores deep learning applications in medical image analysis, emphasizing the significance of explainable AI and discussing techniques to improve interpretability. It also highlights the use of cascaded models for better results. The paper concludes by addressing future directions and challenges in explainable AI for medical image analysis.

The authors' study in [11] emphasizes the importance of explainable models in healthcare predictive analytics. They propose a clinical decision support system for forecasting hospital readmission within 30 days of discharge. By comparing a baseline logistic regression model with an explainable lasso algorithm, known for its variable selection capabilities, the evaluation shows improved performance in the lasso algorithm's area under the ROC curve score. This study provides valuable insights into the effectiveness of the lasso algorithm for this classification task.

[12] introduces NOVA, an innovative annotation tool for analyzing emotional behavior. NOVA's interactive workflow involves active human participation. It utilizes semi-supervised active learning, leveraging machine learning techniques to automatically pre-label data during annotation. Notably, NOVA incorporates state-of-the-art explainable AI (XAI) methods, providing users with confidence values and visual explanations for automatic predictions.

III. METHODOLOGY

A. Datasets

The model was trained on the Brain Tumor Classification (MRI) dataset [13], a robust collection of MRI scans designed specifically for classifying brain tumors. The dataset includes a variety of brain MRI images from patients with different tumor types, such as no tumor, glioma tumor, meningioma tumor, and pituitary tumor, totaling 3,264 files.

B. Data Preprocessing

To ensure uniformity, all images in the dataset were resized to 150 × 150 × 3 dimensions. During training, the order of images was randomized to expedite convergence and prevent the convolutional neural network (CNN) from memorizing the training sequence. Figure 1 showcases sample images representing each tumor type: (A) normal brain without tumors, (B) glioma tumor with irregular boundaries, (C) meningioma tumor originating in the meninges, and (D) pituitary tumor with unique features that present surgical challenges.

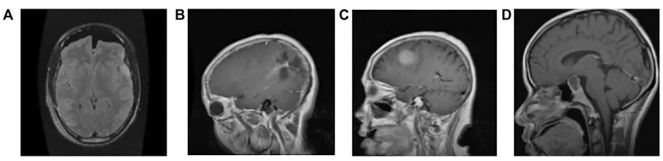


Fig. 1: Sample image data of different types of tumours.

C. Proposed Architecture

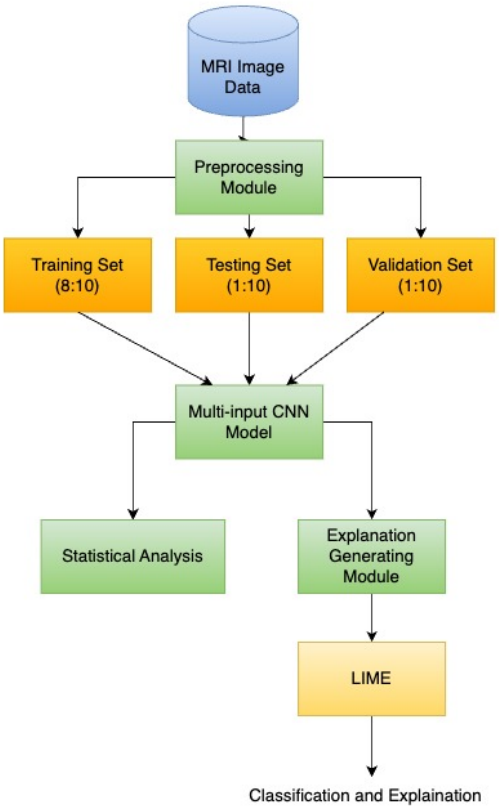


Fig. 2: Block Diagram for Explainable AI for Brain Tumor Detection

The proposed system, illustrated in Figure 2, consists of several stages, including feature extraction, a CNN model, statistical performance metrics, and explanation generation frameworks. To improve accuracy, the CNN model was trained on two copies of the dataset, with an output layer configured to have a 1 × 4 structure. The Adam optimizer, set with default parameters, was used for optimization, and the ReLU and softmax functions were implemented as activation functions. The final CNN model was employed for various tasks, including measuring statistical accuracy and generating explanations through LIME and SHAP methods. For SHAP explanations, a gradient-based approach was utilized, while LIME explanations were derived by computing perturbations.

The system's explainable model employs a dual-input CNN architecture for classification. ReLU activation is applied to all hidden layers, selected for its straightforward computation and its simplicity in determining derivative values, either 0 or 1, based on input positivity. The Adam optimizer with default parameters and sparse categorical cross-entropy as the loss function were used. A kernel size of 3 × 3 was applied in convolutional layers to effectively capture local patterns in the input data.

IV. RESULTS

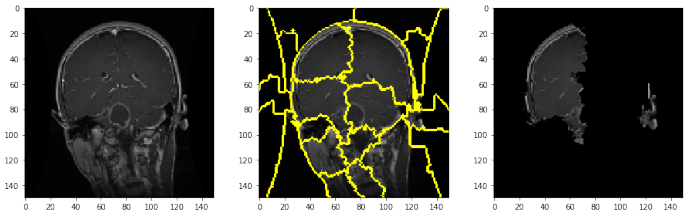
A. CNN Model

The model was trained for 50 epochs, with a monitoring mechanism using min mode and a patience level of three to control overfitting during CNN callbacks. After training, the model made 26 incorrect predictions, achieving a training loss of 0.062. It reached a training accuracy of 97.34% and a final test accuracy of 88.15%, as depicted in Figure 3.

Fig. 3: Training and Validation Accuracy and Loss

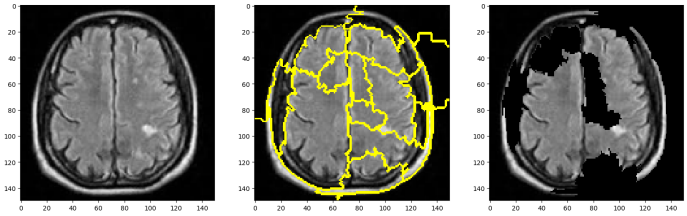
B. LIME

A total of 150 perturbations were applied, generating random values of 1 and 0 to form a matrix with rows as perturbations and columns as superpixels. A superpixel is "active" if it holds a value of 1, and "inactive" if 0. The vector length represents the number of superpixels in each image. Perturbations were applied to the test image based on this vector. Figure 6c shows the final perturbed image, highlighting the regions significant for classification as a meningioma tumor.



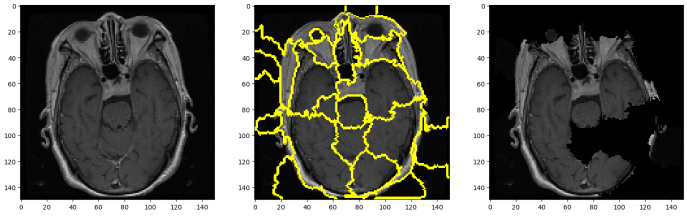
(a) (b) (c)

Fig. 4: Interpretations generated by LIME for meningioma tumour - (a) Sample of meningioma tumour from the test image. (b) Superpixels generated to create perturbations. (c)Final perturbed image showing meningioma tumour



(a) (b) (c)

Fig. 5: Interpretations generated by LIME for no tumour - (a) Sample of no tumour from the test image. (b) Superpixels generated to create perturbations. (c)Final perturbed image showing no tumour



(a) (b) (c)

Fig. 6: Interpretations generated by LIME for Glioma tumour - (a) Sample of Glioma tumour from the test image. (b) Superpixels generated to create perturbations. (c)Final perturbed image showing glioma tumour

V. CONCLUSION AND FUTURE WORK

The model achieved an accuracy of 88.15% on the Brain Tumor Classification (MRI) dataset, demonstrating its capability to accurately identify brain tumors and assist doctors by highlighting critical areas, potentially saving time and lives. Future work involves expanding the dataset to enhance accuracy and accommodate a broader range of tumor types. Additionally, we aim to enhance the system's versatility by enabling it to predict tumors from various medical imaging sources. We also plan to develop a user-friendly interface with comprehensive explanations, allowing non-specialists to understand and benefit from the system. Ultimately, our goal is to make this technology widely accessible and beneficial in medical environments.

REFERENCES

[1] N. Chapatwala, C. N. Paunwala and P. Dalal, 'An Explainable AI approach towards Epileptic Seizure Detection,' *2022 IEEE 19th India Council International Conference (INDICON)*, Kochi, India, 2022, pp. 1-6, doi: 10.1109/INDICON56171.2022.10039982.

[2] T. S. Apon, M. M. Hasan, A. Islam and M. G. R. Alam, 'Demystifying Deep Learning Models for Retinal OCT Disease Classification using Explainable AI,'*2021 IEEE Asia-Pacific Conference on Computer Science and Data Engineering (CSDE)*, Brisbane, Australia, 2021, pp. 1-6, doi: 10.1109/CSDE53843.2021.9718400.

[3] I. S. Gillani, M. Shahzad, A. Mobin, M. R. Munawar, M. U. Awan and M. Asif, 'Explainable AI in Drug Sensitivity Prediction on Cancer Cell Lines,' *2022 International Conference on Emerging Trends in Smart Technologies (ICETST)*, Karachi, Pakistan, 2022, pp. 1-5, doi: 10.1109/ICETST55735.2022.9922931.

[4] Patrick Ferdinand Christ, 'Automatic Liver and Tumor Segmentation of CT and MRI Volumes Using Cascaded Fully Convolutional Neural Networks,'2017 arXiv:1702.05970v2 [cs.CV]

[5] A. Temenos, M. Kaselimi, I. Tzortzis, I. Rallis, A. Doulamis and N. Doulamis, 'Spatio-Temporal Interpretation of The Covid-19 Risk Factors Using Explainable Ai,' *IGARSS 2022 - 2022 IEEE International Geoscience and Remote Sensing Symposium*, Kuala Lumpur, Malaysia, 2022, pp. 7705-7708, doi: 10.1109/IGARSS46834.2022.9884922

[6] M. T. Reza, F. Ahmed, S. Sharar and A. A. Rasel, 'Interpretable Retinal Disease Classification from OCT Images Using Deep Neural Network and Explainable AI,' *2021 International Conference on Electronics, Communications and Information Technology (ICECIT)*, Khulna, Bangladesh, 2021, pp. 1-4, doi: 10.1109/ICECIT54077.2021.9641066.

[7] C. Pan et al., 'Liver Tumor Detection Via A Multi-Scale Intermediate Multi-Modal Fusion Network on MRI Images,'*2021 IEEE International Conference on Image Processing (ICIP)*, Anchorage, AK, USA, 2021, pp. 299-303, doi: 10.1109/ICIP42928.2021.9506237.

[8] U. Pawar, D. O'Shea, S. Rea and R. O'Reilly, 'Explainable AI in Healthcare," *2020 International Conference on Cyber Situational Awareness, Data Analytics and Assessment (CyberSA)*, Dublin, Ireland, 2020, pp. 1-2, doi: 10.1109/CyberSA49311.2020.9139655.

[9] C. Nicodeme, 'Build confidence and acceptance of AI-based decision support systems - Explainable and liable AI," *2020 13th International Conference on Human System Interaction (HSI)*, Tokyo, Japan, 2020, pp. 20-23, doi: 10.1109/HSI49210.2020.9142668.

[10] Saeed Mohagheghi, Amir Hossein Foruzan, 'Developing an explainable deep learning boundary correction method by incorporating cascaded x-Dim models to improve segmentation defects in liver CT images", *Computers in Biology and Medicine,* Volume 140, 2022, 105106, ISSN 0010-4825, doi: 10.1016/j.compbiomed.2021.105106.

[11] A. Vucenovic, O. Ali-Ozkan, C. Ekwempe and O. Eren, 'Explainable AI in Decision Support Systems : A Case Study: Predicting Hospital Readmission Within 30 Days of Discharge,' *2020 IEEE Canadian Conference on Electrical and Computer Engineering (CCECE)*, London, ON, Canada, 2020, pp. 1-4, doi: 10.1109/CCECE47787.2020.9255721.

[12] A. Heimerl, K. Weitz, T. Baur and E. André, 'Unraveling ML Models of Emotion With NOVA: Multi-Level Explainable AI for Non-Experts,' *in IEEE Transactions on Affective Computing,* vol. 13, no. 3, pp. 1155-1167, 1 July-Sept. 2022, doi: 10.1109/TAFFC.2020.3043603.

[13] Sartaj Bhuvaji, Ankita Kadam, Prajakta Bhumkar, Sameer Dedge, amp; Swati Kanchan. (2020).*Brain Tumor Classification (MRI)* [Data set]. Kaggle. https://doi.org/10.34740/KAGGLE/DSV/1183165

[14] https://medium.com/@gauravagarwal_14599/explainableai-understandingtheshaplogic586fcf54c1b9Applications

[15] https://svitla.com/blog/interpretingmachinelearningmodelslimeandshap

[16] https://www.kdnuggets.com/2016/08/introductionlocalinterpretable-modelagnosticexplanationslime.html