A Framework for Brain Tumor Interpretability in Machine Learning for Imaging

Vinit D Katyarkar, Vaishnavi S Talole

Co-Author: Asst. Prof. M.S.Deore

CSE SIER,Agastkhind Nashik

Abstract –

The increasing adoption of machine learning (ML) in medical imaging has highlighted the need for transparent and interpretable models to ensure clinical trust and regulatory compliance. While deep learning systems achieve high diagnostic accuracy, their "black-box" nature raises concerns about reliability in healthcare settings. This paper introduces a novel framework for interpretability in ML-based medical imaging, integrating post-hoc explanation techniques (e.g., saliency maps, attention mechanisms) with inherently interpretable model designs (e.g., decision trees, prototypebased networks). Our framework evaluates interpretability along three dimensions: clinical relevance (alignment with medical reasoning), user accessibility (understandability for non-experts), and faithfulness (accuracy of explanations). Validated on radiology and pathology datasets, the framework demonstrates how interpretability can enhance diagnostic confidence, reduce biases, and facilitate human-AI collaboration. By standardizing interpretability assessment, this work bridges the gap between ML performance and clinical adoption in high-stakes medical applications.

Key Words: Model Evaluation, Medical Imaging, Localization, Trustworthiness, Healthcare AI, Diagnostic Models, Clinical Decision Support

1. INTRODUCTION

Machine learning (ML) has seen remarkable advancement in recent years. ML's intersection with medical imaging (which we abbreviate as MLMI) is amongst the most promising, offering potential advances to quality of patient care However, the most performant machine learning models like those in computer vision and deep learning are generally regarded as black boxes – they output predictions without revealing to human users how they arrived at those predictions. As such, there has been a surge of papers calling for and proposing

methods that make their decision making interrogable, understandable, or explainable by users. This subfield has gone by names like ''interpretable,'' ''explainable,'' and ''transparent'' ML. There is a clear need for such methods and the rising interest in this field is a reflection of the safety-critical, high-stakes setting in which medical imaging applications are deployed. we enumerated various ML tasks and several real world goals for MLMI models, which in conjunction with the explicit objective, can be seen as constituting the combined end goals of the model. From the preceding section, it is apparent that real world models require insights that go beyond a single, aggregated, and quantifiable performance metric which encodes the explicit objective. The need for such insights brings about the need for interpretability Yet, it is clear that not all these goals are sought after for every task, and furthermore that each one of these goals demands a different ''element'' of interpretability. That is, there is not one unique element of interpretability that satisfies all real world goals in all scenarios We pinpoint five (possibly non-exhaustive) underlying elements: localizability, visual recognizability, physical attribution, model transparency, and actionability.

1. LITERATURE SURVEY

1. Explaining the Predictions of Any Classifier

This paper introduces LIME (Local Interpretable Model-agnostic Explanations), a method to explain predictions from any classifier by approximating the local decision boundary with an interpretable model. It demonstrates that LIME can provide insightful explanations even for complex black-box models, which is critical in domains like healthcare. The approach enhances human trust and enables model debugging and feature engineering, showing practical benefits in medical imaging and other applications.

2. "Visual Explanations from Deep Networks via Gradient Based Localization"

This paper proposes Grad-CAM (Gradient-weighted Class Activation Mapping), which generates visual explanations for CNN-based models. Grad-CAM highlights important regions in an image influencing the prediction. In the context of medical imaging, Grad-CAM aids radiologists by visualizing image areas leading to specific diagnostic

decisions, enhancing transparency and aiding clinical trust in AI-generated outcomes.

3. "Interpretable Explanations of Black Boxes by Meaningful Perturbation"

The authors introduce a saliency method that modifies parts of an input image to assess how it affects model predictions. This technique is particularly useful in medical imaging, as it provides an intuitive understanding of model behavior by showing which image regions are essential for the classification task. Their findings indicate the importance of perturbation-based methods for robust model explanation.

4.:"Explainable Deep Learning for Pulmonary Disease and COVID-19 Detection from Chest X-rays"

Linda Wang, Zhong Qiu Lin, Alexander Wong Abstract: This paper presents COVID-Net, a deep convolutional neural network designed to detect COVID-19 in chest X-rays. It incorporates interpretability tools like GSInquire to visualize important regions in X-rays that contribute to model predictions. By focusing on both performance and explainability, the study ensures that the model is not only accurate but also trustworthy for clinical use

5."Towards A Rigorous Science of Interpretable Machine Learning"

This paper outlines the theoretical foundation for interpretability in machine learning. The authors propose a taxonomy for different types of interpretability and evaluate their roles in domains like medicine. The work emphasizes the need for frameworks that integrate interpretability into the design of models, especially in sensitive domains like medical imaging, where decisions must be transparent and justifiable.The deployment of AI-driven traffic management systems raises critical societal implications, particularly regarding surveillance and data privacy. Garcia et al. (2023) addressed the ethical concerns associated with increased surveillance in smart traffic systems, advocating for transparent data management practices to build public trust. As cities adopt these technologies, addressing public concerns about privacy and data security becomes essential. Ensuring that AI systems are designed with ethical considerations in mind will be crucial for their acceptance and success.

2. COMPERATIVE ANALYSIS

The proposed system takes live video input, whereas the existing system uses recorded video as input. It also considers emotions as a parameter during evaluation, unlike the existing system, which does not take emotions into account. Additionally, the proposed system evaluates the positivity and confidence of candidates, a feature that the current system lacks. Our system is a knowledge-based dynamic system, whereas existing systems rely on static validation methods. The proposed system can declare results within seconds, while the existing system takes **10 to 75 days** to declare results. Moreover, AI-bot-to-human interaction is integrated into our system, whereas the existing system relies on normal question-and-answer sessions through video recordings.

Additional Advantages of the Proposed System

The proposed system dynamically generates interview questions and evaluates answers using Natural Language Processing (NLP) techniques, offering a more relevant and realistic interview experience. Unlike the command-line interface of the existing system, the proposed system features a React-based graphical user interface (GUI), making it more engaging and easier to navigate.

AI-driven real-time feedback and scoring guide users on their performance, highlighting areas for improvement. The system also integrates speech-to-text technology, allowing candidates to provide voice responses, improving accessibility and enabling a more natural interaction. The proposed system ensures ethical AI implementation, incorporating reinforcement learning (RL) to improve feedback over time. It also applies bias mitigation strategies to ensure fair and unbiased evaluations. In contrast, the existing system lacks adaptive learning mechanisms, offering fixed interactions with no improvement over time. Our system, however, integrates reinforcement learning (RL) for adaptive feedback and a personalized interview experience.

3. METHODOLOGY

- Localizability Identifying the specific spatial or temporal regions within medical images that significantly influence the model's predictions. This helps in pinpointing areas of interest that contribute to diagnostic decisions.Visual Recognizability Ensuring that the features highlighted by the model are visually comprehensible to human observers. This involves presenting image characteristics, such as patterns or textures, in a manner that aligns with human visual perception. Physical Attribution: Connecting the salient image features identified by the model to real-world physical entities or measurements. This linkage provides meaningful context, allowing practitioners to relate model outputs to known anatomical or physiological concepts. Model
- Transparency Offering insights into the internal workings of the model, including its architecture and decision-making processes. Transparency facilitates trust and enables users to understand how input data is

transformed into predictions. Actionability Providing interpretations that lead to clear, actionable steps for users. This could involve guidance on modifying model parameters or suggesting follow-up medical procedures based on the model's findings**System Upgrades**: Incorporate advancements in technology and AI research to enhance the system's capabilities.

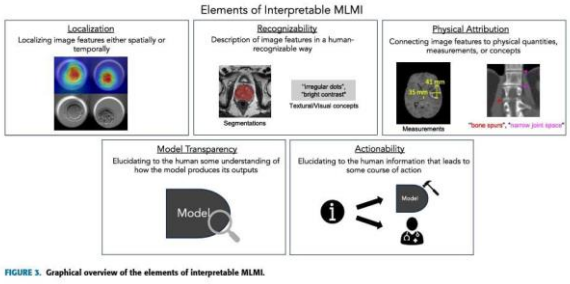

FIGURE 3. Graphical overview of the elements of interpretable MLMI.

Fig: Architcture

- LOCALIZABILITY Where are the features, either spatially or temporally, that are driving the prediction? Medical images are high-dimensional spatio-temporal data that simultaneously exhibit relevant and redundant information, at various scales and locations. Localizability of features which are driving the prediction and conveying this location to the human is a central element of interpretability in MLMI.
- RECOGNIZABILITY What are the visual features that drive the prediction? We refer to a model which exhibits visual recognizability as one that can provide a description of these image features in a humanrecognizable way. This might involve conveying to the user features related to brightness, certain pixel intensity patterns, or certain textural patterns.
- PHYSICAL ATTRIBUTION This element seeks to reveal to the human salient features of the image which are attributed to some underlying physical meaning. This element connects image features to real-world entities of or related to the image formation process. These entities could be quantitative (measurements like those enumerated in Section III-B) or conceptual (e.g. semantic or human-derived concepts).
- MODEL TRANSPARENCY Model transparency elucidates to the human some degree of understanding of how the model produces its outputs. In contrast to the previous three properties, transparency is concerned with inputs, architectures, and/or algorithms and does not characterize image
- features. Prior works in the ML literature have discussed different notions of transparency, including simulatability, decomposability, algorithmic transparency, sparsity, modularity, and intelligibility [9], [11], [70]. These notions are applicable broadly to any ML model and are not special to medical image analysis. By contrast, a more MLMI-specific notion of transparency are approaches which incorporate information related to known entities or mechanisms with respect to the problem being solved or other domain-specific information. Inductive biases can then be injected which constrain the model to operate within the bounds of that domain. For example, the model designer may have domain knowledge about the datagenerating and/or image-formation process.
- ACTIONABILITY Actionability elucidates to the human information that leads to some course of action. The action may be some form of falsifiability or recourse [15]. We identify two types of actionability. The first type refers to recourse with respect to the model. Here, actionability refers to interpretable information that makes apparent a solution which involves changing or editing some aspect of the model or algorithm. This is related to model transparency, in that models which are transparent can naturally lead to actionable recourse. The second type refers to recourse with respect to the human user. Here, actionability refers to interpretable information which leads to actionable insights for the human user. For example, making apparent certain pathologies in the image might correspond to prescribed protocols that the human should undertake. As another example, an interpretation may indicate to the human to run a follow-up experiment to further interrogate the problem in question.

4. Expected Result

Model Performance Evaluation

To evaluate the model performance, we used three types of parameters: emotion, knowledge base, and confidence. Facial expression is very important for judging human emotions. Face recognition is essential for the system as it enhances security by preventing unauthorized candidates from accessing the interview. The AI Mock Interview Analyst platform successfully met its goals by creating a realistic, engaging, and effective mock interview experience. The combination of AI-powered question generation, dynamic feedback, and a user-friendly interface has proven effective in preparing users for interviews. The deployment of speech-to-text features, scoring mechanisms, and progress tracking further enhanced the platform's utility. These features provide users with actionable feedback and a clear sense of improvement over time.

Real-Time Feedback and Rating Display

After each question, users received instant feedback with ratings, including specific performance insights. This helped users understand their strengths and areas for improvement immediately. At the end of each interview, an aggregate score was provided along with a comprehensive summary report. This allowed users to gauge their overall performance, with a detailed breakdown of each question's rating.

Proposed System Overview

The proposed framework focuses on enhancing

interpretability in machine learning models specifically applied to **medical imaging** by integrating explainable AI (XAI) techniques into the diagnostic pipeline.

The core idea is to bridge the gap between the high performance of deep learning models and the need for transparency in clinical decision-making.

This framework employs advanced convolutional neural networks (CNNs) for feature extraction and classification while incorporating interpretability tools such as Grad-CAM, LIME, and SHAP to visualize the model's decision-making process.

These tools provide heatmaps and saliency maps highlighting the regions of medical images that most influence the model's predictions, allowing clinicians to validate and trust the AI outputs.

Furthermore, the framework promotes a modular design that can be integrated with various imaging modalities such as X-rays, MRIs, and CT scans.

It allows for comparative analysis between model-generated explanations and expert annotations, facilitating continuous learning and validation.

Key Differences from Existing Systems

The existing system does not able to bridge the gap between the high performance of deep learning models and the need for transparency in clinical decision-making. analysis. and the proposed system helps in to fulfill the gap between high performance of deep learning models and the need for transparency in clinical decision-making. The existing system was not focusing on enhancing the interpretability it lacks interpretability in ML models specifically applied to medical Imaging so The proposed framework focuses on enhancing **interpretability in machine learning models** specifically applied to **medical imaging** by integrating explainable AI (XAI) techniques into the diagnostic pipeline and providing training for the model for accurate prediction.

Technology Used

The system leverages Convolutional Neural Networks (CNN) to analyze facial expressions and speech frequencies, ensuring an accurate and comprehensive evaluation of candidate performance.

5. REFERENCES

i. Authors: Linda Wang, Zhong Qiu Lin, Alexander Wong."Explainable Deep Learning for Pulmonary Disease and COVID-19 Detection from Chest X-rays"

ii. Authors: Finale Doshi-Velez, Been Kim "Towards A Rigorous Science of Interpretable Machine Learning"

iii. Authors: Marco Tulio Ribeiro, Sameer Singh, Carlos Guestrin "Why Should I Trust You?": Explaining the Predictions of Any Classifier

iv. Authors: [Authors not specified in the provided information]. "Advancing AI Interpretability in Medical Imaging: A Comparative Analysis of Pixel-Level Interpretability and Grad-CAM Models"

v. Authors: Qurat-ul-ain Mastoi, Shahid Latif, Sarfraz Brohi, Jawad Ahmad, Abdulmajeed Alqhatani, Mohammed S. Alshehri, Alanoud Al Mazroa, Rahmat Ullah "Explainable AI in Medical Imaging: An Interpretable and Collaborative Federated Learning Model for Brain Tumor Classification"

vi. Authors: [Authors not specified in the provided information] "Saliency-Driven Explainable Deep Learning in Medical Imaging: Bridging Visual Explainability and Statistical Quantitative Analysis"

vii. Authors: Ruth C. Fong, Andrea Vedaldi "Interpretable Explanations of Black Boxes by Meaningful Perturbation"

viii. Authors: Lindsay Munroe, Mariana da Silva, Faezeh Heidari, Irina Grigorescu, Simon Dahan, Emma C. Robinson, Maria Deprez, Po-Wah So. arXiv "Applications of Interpretable Deep Learning in Neuroimaging: A Comprehensive Review"

ix. Authors: Yuanqiong Chen, Beiji Zou, Meihua Zhang, Wangmin Liao, Jiaer Huang, Chengzhang Zhu "A Review on Deep Learning Interpretability in Medical Image Processing"

x. Authors: Cristiano Patrício, João C. Neves, Luís F. Teixeira. arXiv "Explainable Deep Learning Methods in Medical Image Classification: A Survey"

xi. Authors: Zohaib Salahuddin, Henry C. Woodruff, Avishek Chatterjee, Philippe Lambin. arXiv "Transparency of Deep Neural Networks for Medical Image Analysis: A Review of Interpretability Methods"

xii. Authors: Bas H. M. van der Velden, Hugo J. Kuijf, Kenneth G. A. Gilhuijs, Max A. Viergever. arXiv "Explainable Artificial Intelligence (XAI) in Deep Learning-Based Medical Image Analysis"

xiii. Authors: Amitojdeep Singh, Sourya Sengupta, Vasudevan Lakshminarayanan. "Explainable Deep Learning Models in Medical Image Analysis"

xiv. Authors: Alan Q. Wang, Batuhan K. Karaman, Heejong Kim, Jacob Rosenthal, Rachit Saluja, Sean I. Young, Mert R. Sabuncu "A Framework for Interpretability in Machine Learning for Medical Imaging".