International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING
ISSN:2147-6799 www.ijisae.org Original Research Paper

Multi-Modal Explainability Evaluation for Brain Tumor Segmentation: Metrics MSFI

Maria Nancy A.[1], K. Sathyarajasekaran[2*]

Submitted: 02/12/2023 **Revised**: 10/01/2024 **Accepted**: 26/01/2024

Abstract: The significance of interpretability in artificial intelligence (AI) models is growing within the healthcare sector, driven by advancements in medical imaging technology. These developments enhance our ability to recognize and understand intricate biomedical occurrences. As medical imaging technology progresses, the need for interpretable AI models becomes more critical in ensuring trust, accountability, and acceptance among healthcare professionals. In this context, the Multi-Modal Specific Feature Importance (MSFI) metric emerges as a crucial tool for evaluating the effectiveness of eXplainable Artificial Intelligence (XAI) models, specifically Grad-CAM, in multi-modal medical imaging tasks. The MSFI metric addresses the intricacies of interpreting decisions made by AI models when presented with multi-modal medical images. Clear and detailed explanations are essential for ensuring a thorough comprehension and fostering trust in the decision-making process. This is particularly crucial as these visuals communicate diverse clinical information pertaining to the same underlying biomedical reality. The metric aims to assess how well heat-maps or feature attribution maps elucidate these decisions. The evaluation process using the MSFI metric is a comprehensive approach that combines computational methods with clinician user studies. For assessing the challenging brain tumor segmentation task clinically, the MSFI metric serves as a valuable tool. This metric gauges the correlation between the model prediction and the plausibility measure from various explainable artificial intelligence (XAI) approaches. In the selection and development of XAI algorithms tailored to meet clinical requirements for multi-modal explanation, the MSFI metric proves to be a valuable resource. By focusing on addressing the interpretability of modality-specific features, this metric provides a framework for refining and advancing XAI models in the realm of medical imaging. The MSFI measure offers a robust evaluation framework that aids in comprehending the performance of AI models in the intricate realm of multi-modal medical imaging, particularly in the context of brain tumor segmentation diagnosis.

Keywords: *Explainable AI, Multi Modal Specific Feature Importance (MSFI), Modality Importance(MI).*

1. Introduction

Numerous critical decisions in healthcare, finance, and situations involving life-or-death consequences depend on AI systems. Consequently, the need for explainable AI (XAI) is imperative to establish public trust and confidence in these systems [1]. Explainable AI (XAI) involves elucidating the rationale behind the decisions, recommendations, or predictions made by an AI system. This capability empowers human users to comprehend and trust the results and outcomes generated by machine learning algorithms [2]. The primary value of Explainable AI (XAI) lies in its ability to minimize the chances of ethical violations, biases, and legal complications. This is achieved through ensuring compliance, transparency, and accountability in AI decision-making [3]. XAI also helps promote end-user trust, model auditability, and productive use of AI, while mitigating compliance, legal, security, and reputational risks of production AI [4]. Researchers can adopt a more conscientious approach to AI development with the assistance of XAI, which elucidates AI decisions and behaviors in a manner comprehensible to humans. This leads to the creation of AI systems that are more trustworthy and responsible [5].

Evaluating the quality and effectiveness of Explainable Artificial Intelligence (XAI) is essential for various reasons, including trust, accountability, and ethics, especially in domains such as healthcare, finance, education, and security. There are several dimensions to evaluate XAI, including user, task, model, and data [6]. Some key reasons for evaluating XAI models are: Usability: Assessing the ease of use and understanding of the XAI model for human users, considering their background, goals, preferences, and level of expertise. Accuracy: Ensuring that the XAI model provides accurate and reliable explanations of the decision-making process, helping users understand the reasoning behind the model's output [7]. Fairness and Transparency: Evaluating the fairness and transparency of the XAI model, which can help identify potential biases and ensure that the model is not discriminatory. Model Performance: Investigating the impact of the XAI model on the performance of the underlying machine learning algorithm, and identifying areas for improvement. Compliance and Legal: Ensuring that the XAI model adheres to regulatory standards and

[1]*Research Scholar, School of Computer Science and Engineering, Vellore Institute of Technology, Chennai, Tamil Nadu, India.*
[1]*Email: marianancy.a2017@vitstudent.ac.in*
[2*]*Associate Professor, School of Computer Science and Engineering,, Vellore Institute of Technology, Chennai, Tamil Nadu, India*
** Corresponding Author Email: sathyarajasekaran.k@vit.ac.in*

legal requirements, and can be used as evidence in case of disputes or challenges. Trust and Accountability: Building trust and confidence in the AI system by providing clear and understandable explanations of its actions and decisions. Encourage AI explainability and assess the business implications of implementing such algorithms through consistent monitoring and management of models. This methodology in AI development is commonly referred to as responsible AI development. To evaluate XAI models, it is crucial to use a combination of qualitative and quantitative methods, such as application-grounded metrics, user studies, and comparative analysis with baseline models [8]. By doing so, organizations can ensure the quality and effectiveness of their XAI models, leading to more trustworthy and responsible AI systems.

In this study, it is recommended to conduct a systematic evaluation grounded in clinical requirements using computational tools and clinician user studies to analyze the MSFI metric. The objective of this evaluation is to assess the effectiveness of heatmaps and feature attribution maps in providing decision-supportive context for multi-modal medical images. This is particularly relevant in cases where different imaging modalities present diverse clinical data associated with the same biological event. The MSFI metric measures both the model prediction for brain tumor segmentation and the plausibility measure of all XAI approaches. To calculate it, a modality-specific feature importance (MI) is adjusted to a weighted sum of all heatmap values within the feature localization mask for each modality [9]. The evaluation results, along with the MSFI metric, can guide the development and selection of XAI algorithms to meet clinicians' needs for a multi-modal explanation. Ultimately, obtaining input from neurosurgeons is crucial to determining the overall understandability.

2. Related Work

2.1 How XAI works in Medical Imaging:

XAI techniques for image-based deep learning models can be integrated to make the model's decisions interpretable. Some of the popular techniques include: i)Saliency Maps: These visualizations indicate the most relevant parts of the input images for making predictions. They contribute to understanding how the model focuses on specific regions within the image [10]. ii) Grad-CAM (Gradient-weighted Class Activation Mapping): Grad-CAM generates a class-discriminative localization map by analyzing the gradient information related to the target class in the input image. This map serves as a visual representation of the model's prediction, highlighting areas of significance. iii) Local Interpretable Model-agnostic Explanations (LIME): LIME provides explanations for specific predictions at a local level, offering insights into how the model arrives at its decisions [11]. iv) SHAP (SHapley Additive exPlanations): SHAP assigns feature importance to pixels in an image, helping in understanding the impact of individual features on the model's prediction[12]. These XAI techniques can be applied to image-based deep learning models to provide interpretability and explainability, enabling clinicians and other end-users to trust the model's recommendations and understand its decision-making process. Figure 1 shows that the Steps of XAI based model Evaluation.

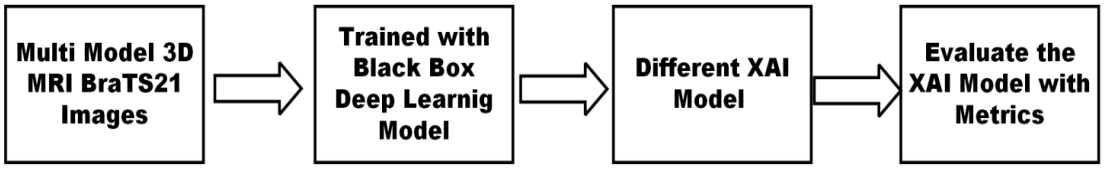

Fig 1: Steps of XAI based model Evaluation

Multi Model Images:

Multi-modal 3D MRI refers to the use of multiple imaging modalities, such as T1-weighted, T2-weighted, and FLAIR, in three-dimensional magnetic resonance imaging for the comprehensive assessment of brain structures and pathologies. This approach allows for the acquisition of complementary information from different imaging sequences, which can improve the accuracy of diagnosis and treatment planning for various neurological conditions, including brain tumors, Alzheimer's disease, and stroke[13]. The assessment of brain tumor segmentation in 3D multi-modal MRI data is commonly conducted using the BRATS (Multimodal Brain Tumor Segmentation) challenge, serving as a pivotal platform. This dataset comprises three-dimensional magnetic resonance imaging (MRI) images acquired through four distinct modalities: first-pass (T1), post-contrast T1-weighted (T1ce), second-pass (T2), and third-pass (T2-FLAIR) fluid-attenuated inversion recovery. The images are carefully aligned and resampled to achieve an isotropic resolution of 1x1x1 mm.

The outcomes of these assessments showcase promising advancements in the segmentation of brain tumors within multi-modal MRI images. The BRATS challenge consistently proves to be an invaluable resource, facilitating the refinement and evaluation of cutting-edge segmentation methodologies tailored for the comprehensive analysis of brain tumors.Figure2 shows Sample Multi Model MRI images with different modalities like T1, T2, T1ce, T2-Flair and Mask.

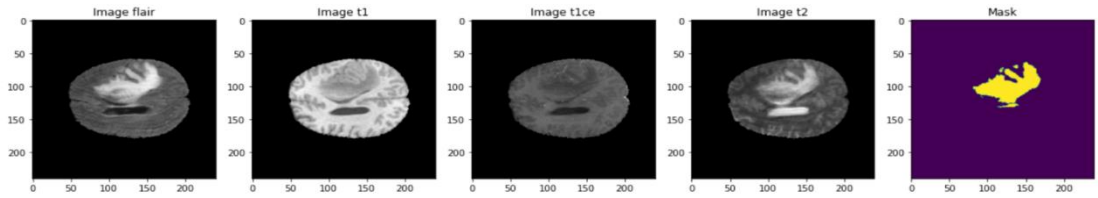

Fig 2: Sample Multi Model MRI images with different modalities like T1, T2, T1ce, T2-Flair and Mask

Modality Importance (MI)

The concept of Modality Importance, as determined by Shapley values, draws inspiration from cooperative game theory, providing an equitable method for distributing the cumulative contributions of each modality within a set. Shapley values are particularly advantageous due to their commendable properties, which include efficiency, symmetry, linearity, and marginalism. Within cooperative game theory, Shapley values offer a fair mechanism to allocate a payoff among a group of players, considering their individual marginal contributions to various coalitions. In the realm of Modality Importance for model predictions, each modality is analogous to a player, and the Shapley value serves as a metric for its equitable contribution to the overall model performance. The Modality Shapley value, denoted as $\varphi_{m.}$ The symbol "m" represents the authentic Modality Importance score. Its calculation involves the use of the Shapley value formula:

$$\varphi_m(v) = \sum_{c \subseteq \mathcal{M} \setminus \{m\}} \frac{|c|!(M-|c|-1)!}{M!} (v(c \cup \{m\}) - v(c)) \quad (1)$$

In this context, "v" denotes the performance metric specific to a modality, and $M\{\mathrm{m}\}$ encompasses all subsets of modalities that do not include modality "m." In our evaluation, "v" is defined as the test set accuracy of the prediction model. To assess the performance of the attributed modalities, we assign a value of zero to all characteristics in a modality that is not part of the subset. The resulting modality Shapley value is denoted as φmod. We also explored an alternative approach where we sampled from areas outside of lesions and used those results to replace zeros with non-zero values for an ablated modality. However, as the rank and magnitude of the Shapley values derived using this method were similar across modalities, we chose the simpler zero replacement setting.

Following the generation and post-processing of saliency maps for each method, a comparison is made between their modality importance values and the previously defined ground truth. Subsequently, the estimated modality importance is calculated by summing up all positive values in the saliency map associated with that modality. To assess the consistency in rating the relevance of modalities between the ground truth and the estimated values, the MI Correlation is measured. In this evaluation, the test set and Kendall's Tau-b correlation are utilized.Figure 3 showsTechniques for Computational Evaluation.

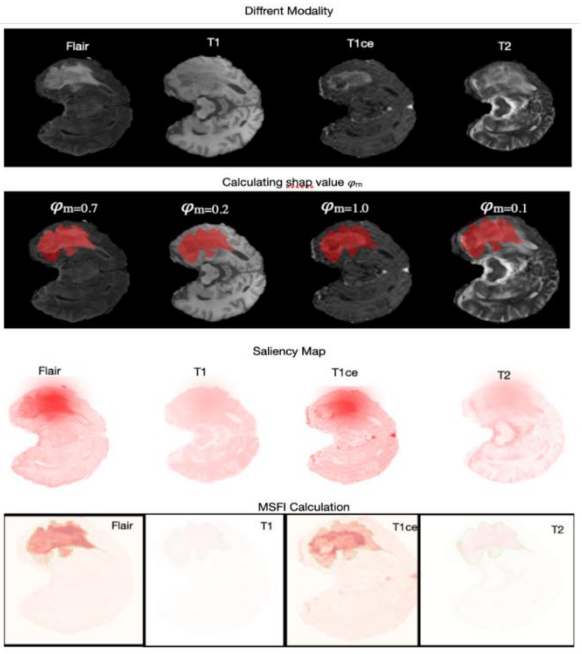

Fig 3: Techniques for Computational Evaluation

The Modality-Specific Feature Importance (MSFI)

This metric offers a framework to evaluate the comprehensibility of AI models in the context of medical imaging tasks that encompass multiple modalities. The MSFI metric evaluates how well the model's prediction aligns with the plausibility measure of various XAI approaches in the brain tumor segmentation task. It is calculated by summing the share of heatmap values inside the feature localization mask for each modality. Each modality's share is weighted by the normalized value of the modality-specific feature importance (MI).

A comprehensive assessment, grounded in clinical requirements and utilizing computational methodologies alongside clinician user studies, is employed in evaluating the MSFI metric. The results of this evaluation can guide the development and selection of XAI algorithms to meet the clinical needs for multi-modal explanation. The MSFI metric provides a framework for evaluating the explainability of the Grad-CAM XAI model in the context of multi-modal medical imaging tasks, with a focus on addressing clinical requirements and the interpretation of modality-specific features. The MSFI metric is a valuable tool for assessing the reliability, generalizability, and interpretability of XAI algorithms in the context of multi-modal medical imaging tasks. To meet clinicians' requirements for a multi-modal explanation, the results of the evaluation and the MSFI metric can guide the development and selection of XAI algorithms.

$$\hat{MSFI} = \sum_m \varphi_m \frac{\sum_i \mathbb{1}(L_m^i > 0) \odot S_m^i}{\sum_i S_m^i} \quad (2)$$

$$MSFI = \frac{\hat{MSFI}}{\sum_m \varphi_m} \quad (3)$$

In the assessment of Modality-Specific Feature Importance (MSFI), data from feature localization masks or bounding boxes is integrated with Modality Importance (MI). For each modality, MSFI represents the proportion of saliency map values located within the corresponding legitimate feature localization mask. The normalized Modality Importance value, denoted as φ_m, is then applied as an additional weighting factor.

For a given modality "m," let "S" denote its saliency map, with "i" representing the spatial position. The ground truth localization masks or bounding boxes for modality "m" are represented as "L_m" when describing the spatial region of the feature where$L_m^i > 0$. The indicator function, denoted as $\mathbb{1}$, utilizes the feature mask to selectively extract the saliency map values. φ_m is a normalized Modality Importance value ranging from 0 to 1 for modality "m." The normalized form of MSFI, represented by a metric $\hat{MSFI}$ that can assume values between 0 and 1, is derived from an unnormalized metric.

In our evaluation, we adopt $S_m^i = \mathbb{1}(S_m^i > 0) \odot \hat{S}_m^i$, where S_m^i is the saliency map containing only positive values. A higher MSFI score, as depicted in Fig. 1, signifies a saliency map that effectively captures crucial modalities and their localized features. Notably, MSFI differs from conventional metrics like Intersection over Union (IoU) in its reduced reliance on either saliency map signal intensity or the area of the ground truth localization mask, rendering it a robust metric. In the subsequent sections, we detail our evaluation experiments applying MSFI to a real dataset (BraTS)

During our review, we employed the saliency map containing exclusively positive values. we adopt $S_m^i = \mathbb{1}(S_m^i > 0) \odot \hat{S}_m^i$, where S_m^i illustrates that a saliency map successfully capturing crucial modalities and their localized features results in a higher MSFI score. In contrast to more conventional metrics like Intersection over Union (IoU), MSFI stands out as a robust metric as it is less reliant on the signal intensity of the saliency map or the area of the ground truth localization mask. The results of our evaluation studies using MSFI on a real-world dataset (BraTS) are detailed in the following sections.

3. Results and Discussion

In the BraTS Segmentation challenge, where XAI approaches were evaluated for their fidelity to the model decision process at both modality and feature levels, most MSFI scores fell within the lower range. Notably, no XAI approach achieved an average MSFI score higher than 0.5. Among the tested XAI methods, only Guided GradCAM exhibited statistically significant outperformance ($p < 0.01$). This conclusion was reached through a post-hoc Nemenyi test following a significant Friedman test ($\chi2(15) = 1540.6$, $p < 0.001$). In summary, the synthetic data experiment indicated that Guided GradCAM was the most effective XAI method for the glioma challenge. Furthermore, there was no statistically significant relationship between the rankings of synthetic data MSFI and the correlation between MI.

Table 1: XAI algorithm

XAI Model	MI Correlation[0-1]	MSFI Coorelation[0-1]	MI Correlation[0-1]	MSFI Coorelation[0-1]
	VGG16 based Brain Tumor Segmentation		3D Unet based Brain Tumor Segmentation Using BraTs2021	

	Using BraTs2021 Dataset		Dataset	
MACE	0.16 ± 0.11	0.04 ± 0.02	NaN	NaN
Gradient Shap	0.18 ± 0.12	0.22 ± 0.19	0.46 ± 0.31	0.22 ± 0.23
Lime	0.51 ± 0.08	0.15 ± 0.07	0.34 ± 0.42	0.08 ± 0.08
GradCAM	***0.54 ± 0.12**	***0.43 ± 0.24**	0.53 ± 0.27	0.22 ± 0.25
Guided GradCAM	***0.53 ± 0.09**	***0.42 ± 0.29**	***0.63 ± 0.31**	***0.49 ± 0.23**
Smooth Grad	0.31 ± 0.10	0.02 ± 0.13	0.49 ± 0.23	0.10 ± 0.10
GradCAM++	0.36 ± 0.11	0.08 ± 0.02	0.35 ± 0.19	0.03 ± 0.02

Table 1: Table given for each XAI algorithm, the mean ± standard deviation is displayed in the table for two evaluation metrics: MI correlation, and MSFI for BraTs2021 Dataset. The range of each metric is stated. A higher number is preferable for all metrics. The top three scores on a given metric are bolded, and a ∗ indicates that the XAI algorithm outperformed the others by a significant margin. NaN means the results were not a number.

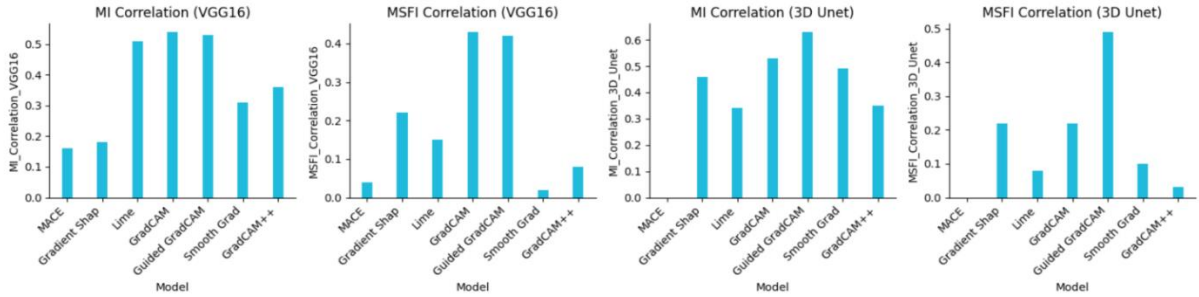

Fig 4: Chart representation of MI correlation and MSFI correlation

The evaluation results for explainable Artificial Intelligence algorithms utilizing VGG16 and Unet models show the mean and standard deviation for MI correlation and MSFI evaluation metrics on the BraTs2021 1250 cases training 3D Dataset and 220 cases test 3D dataset shown in Figure 4, which contain 4 different modalities for every cases. The use of VGG16 and Unet models for these XAI algorithms allows for a comprehensive evaluation of their effectiveness in interpreting the underlying decision-making processes of the neural networks. Incorporating these models in the evaluation allows for a thorough analysis of the explainability and interpretability of the neural network's predictions.

The current trend in artificial intelligence research aims to attain high performance and gain insights into the decision-making processes of machine learning models. This comprehensive review aligns with this direction. In domains such as healthcare, where interpretability of AI algorithms raises significant ethical and legal concerns, it is crucial to understand how machine learning models make decisions.

User Case Study:

Neurosurgeons play a crucial role in the medical field, where AI-based clinical decisions have a significant impact on patient treatment options and outcomes. To understand the concepts of Explainable Artificial Intelligence, neurosurgeons need to be familiar with the integration of AI in surgery and its potential implications for patient care.

This understanding helps neurosurgeons in optimizing surgical paths for preoperative patient images, diagnosing diseases, improving diagnostic efficiency, developing treatment algorithms, making accurate clinical decisions during surgical interventions, and handling repetitive work processes without the risk of burnout.

In the field of medical imaging, the development of methods for dissecting the internal workings of machine learning models is crucial. Post hoc XAI methods, such as heatmap explanations, can provide valuable insights by highlighting important regions in the input image that influence the model's decision [14,15]. These heatmap explanations can help clinicians understand which areas of

the brain MRI are more relevant for classification, aiding in the diagnosis and treatment of diseases. XAI heatmaps can enhance the interpretability and transparency of deep learning models in medical imaging. Received feedback from the neurosurgeon. Feedback from domain experts shown in a stacked bar chart Figure 5. In that chart clearly shows that Guided GradCam got better result of understanding.[16-18]

4. Conclusion

In the exploration of explainability in Brain Tumor Segmentation through multi-modal analysis, the assessment involved three key metrics: MSFI (Modality-Specific Feature Importance), MI (Mutual Information), and Understandability. The study aimed to shed light on the interpretability and effectiveness of explainable Artificial Intelligence (XAI) models in the complex task of brain tumor segmentation. The MSFI metric played a crucial role in evaluating the relevance of each modality in contributing to the segmentation outcomes. By quantifying modality-specific feature importance, the study provided insights into which imaging modalities had a more significant impact on the segmentation results. This not only enhances our understanding of the underlying mechanisms but also guides the refinement of future segmentation models. Mutual Information (MI) emerged as another pivotal metric, measuring the interdependence between the predicted segmentation and ground truth across multiple modalities.

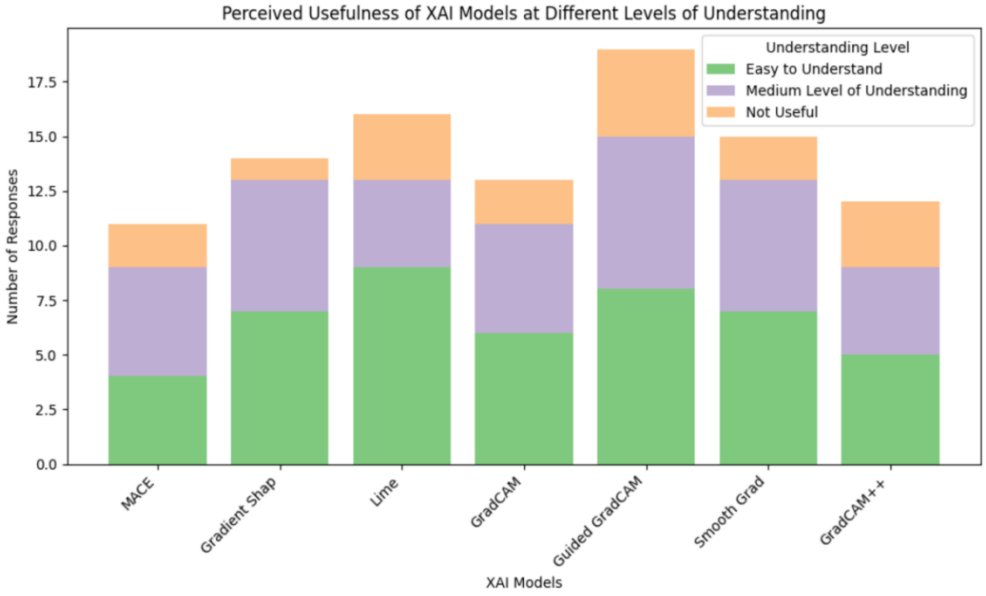

Fig 5: Stacked Bar chart about the feedback given by domain expertise

This metric aimed to quantify the consistency and alignment of model predictions with the actual anatomical structures, providing a comprehensive evaluation of the segmentation accuracy. The evaluation also considered the Understandability metric, recognizing the importance of human interpretability in medical imaging applications. The study assessed how well the XAI models conveyed the segmentation results to human experts, acknowledging the critical role of intuitive and transparent explanations in fostering trust and adoption within the medical community. In conclusion, the multi-modal explainability evaluation utilizing MSFI, MI, and Understandability metrics offers a holistic perspective on the performance and interpretability of XAI models in the challenging domain of brain tumor segmentation. By comprehensively assessing modality-specific contributions, alignment with ground truth, and human interpretability, the study contributes valuable insights for the advancement of explainable AI in medical image analysis, fostering trust and understanding among healthcare practitioners.

References

[1] M. Förster, P. Hühn, M. Klier, and K. Kluge, 2021. Capturing users' reality: A novel approach to generate coherent counterfactual explanations.

[2] A.M. Antoniadi, Y. Du, Y. Guendouz, L. Wei, C. Mazo, B.A. Becker, and C. Mooney, 2021. Current challenges and future opportunities for XAI in machine learning-based clinical decision support systems: a systematic review. *Applied Sciences*, *11*(11), p.5088.

[3] M. Ebers, 2020. Regulating Explainable AI in the European Union. An Overview of the Current Legal Framework (s). *An Overview of the Current Legal Framework (s)(August 9, 2021). Liane*

Colonna/Stanley Greenstein (eds.), Nordic Yearbook of Law and Informatics.

[4] K. Fiok, F.V. Farahani, W. Karwowski, and T. Ahram, 2022. Explainable artificial intelligence for education and training. *The Journal of Defense Modeling and Simulation*, *19*(2), pp.133-144.

[5] A. Sheth, M. Gaur, K. Roy, and K. Faldu, 2021. Knowledge-intensive language understanding for explainable ai. *IEEE Internet Computing*, *25*(5), pp.19-24.

[6] G. Elkhawaga, O. Elzeki, M. Abuelkheir, and M. Reichert, 2023. Evaluating Explainable Artificial Intelligence Methods Based on Feature Elimination: A Functionality-Grounded Approach. *Electronics*, *12*(7), p.1670.

[7] G. Vilone, and L. Longo, 2021. Notions of explainability and evaluation approaches for explainable artificial intelligence. *Information Fusion*, *76*, pp.89-106.

[8] A.J. Johs, D.E. Agosto, and R.O. Weber, 2020. Qualitative investigation in explainable artificial intelligence: A bit more insight from social science. *arXiv preprint arXiv:2011.07130*.

[9] W. Jin, X. Li, M. Fatehi, and G. Hamarneh, 2023. Guidelines and evaluation of clinical explainable AI in medical image analysis. *Medical Image Analysis*, *84*, p.102684.

[10] C. Patrício, J.C. Neves, and L.F. Teixeira, 2023. Explainable Deep Learning Methods in Medical Image Classification: A Survey. *ACM Computing Surveys*, *56*(4), pp.1-41.

[11] A. Chaddad, J. Peng, J. Xu, and A. Bouridane, 2023. Survey of explainable AI techniques in healthcare. *Sensors*, *23*(2), p.634.

[12] B.H. Van der Velden, H.J. Kuijf, K.G. Gilhuijs, and M.A. Viergever, 2022. Explainable artificial intelligence (XAI) in deep learning-based medical image analysis. *Medical Image Analysis*, *79*, p.102470.

[13] D. Nie, J. Lu, H. Zhang, E. Adeli, J. Wang, Z. Yu, L. Liu, Q. Wang, J. Wu, and D. Shen, 2019. Multi-channel 3D deep feature learning for survival time prediction of brain tumor patients using multi-modal neuroimages. *Scientific reports*, *9*(1), p.1103.

[14] X. Hou, D. Yang, D. Li, M. Liu, Y. Zhou, and M. Shi, 2020. A new simple brain segmentation method for extracerebral intracranial tumors. *PLoS One*, *15*(4), p.e0230754.

[15] S. Arndt, C. Turvey, and N.C. Andreasen, 1999. Correlating and predicting psychiatric symptom ratings: Spearmans r versus Kendalls tau correlation. *Journal of psychiatric research*, *33*(2), pp.97-104.

[16] H.W. Loh, C.P. Ooi, S. Seoni, P.D. Barua, F. Molinari, and U.R. Acharya 2022. Application of explainable artificial intelligence for healthcare: A systematic review of the last decade (2011–2022). *Computer Methods and Programs in Biomedicine*, p.107161.

[17] P.N. Srinivasu, N. Sandhya, R.H. Jhaveri, and R. Raut, 2022. From blackbox to explainable AI in healthcare: existing tools and case studies. *Mobile Information Systems*, *2022*, pp.1-20.

[18] A. Adadi and M. Berrada, 2020. Explainable AI for healthcare: from black box to interpretable models. In *Embedded Systems and Artificial Intelligence: Proceedings of ESAI 2019, Fez, Morocco* (pp. 327-337). Springer Singapore.