Chapter 2: Literature Review

Illuminating the black box: an explainable AI framework for brain tumor segmentation using volumetric data interpretation in 3D CNNs

Yogi Amitkumar Patel U2536809

Table of contents

1 Introduction

1.1 Background and Motivation

In modern oncology, properly identifying and managing brain tumours continues to be a major issue. Due to its simplicity and excellent volumetric resolution, magnetic resonance imaging (MRI) has become the highest standard for analysis (Menze et al., 2015). However, there is a substantial barrier in the interpretation of this data. At the moment, the process of defining the boundaries of tumors is mostly done by hand. Subjectivity and labour intensity are introduced by this reliance on human interpretation, which may not be sustainable for scaling healthcare solutions (Iftikhar et al., 2025). Automated technologies that can match or surpass human precision without the corresponding time restrictions are therefore desperately needed in medical applications.

1.2 The "Black Box" Paradox

The response to this clinical need has been the rapid adoption of Deep Learning (DL) technologies, particularly 3D Convolutional Neural Networks (CNNs) like the 3D U-Net. These models have demonstrated "superior performance" in handling volumetric data compared to traditional methods (Iftikhar et al., 2025; Bhati et al., 2024). However, the integration of these advanced models has introduced a new, critical issue: the **"black box"** problem.

While these models are highly accurate, their internal decision-making processes are opaque (Neri et al., 2023).**This creates a paradox in medical AI**: as models become more complex and accurate, they often become less interpretable to the clinicians using them. In a high-stakes environment like neurosurgery, a lack of transparency creates a "trust gap" (Wen et al., 2025). It is not enough for a model to simply output a segmentation mask; clinicians require the ability to verify why specific voxels were classified as tumour tissue. Therefore, the lack of interpretability is no longer just a technical issue, but a barrier to ethical and safe clinical deployment (Neri et al., 2023).

1.3 Research Aims and Scope of Literature Review

To address this disconnect between model accuracy and clinical trust, this literature review aims to critically evaluate the intersection of 3D Deep Learning, Explainable AI (XAI), and Immersive Visualisation. While these fields are often studied in isolation, this review argues that a unified framework is necessary to fully illuminate the "black box" of 3D CNNs.

The review is structured to build this argument logically:

Section 2.0 establishes the technical capabilities of modern 3D CNN architectures for segmentation.

Section 3.0 critically analyses current XAI methods (such as Grad-CAM and LIME), evaluating their limitations when applied to complex volumetric data.

Section 4.0 explores the potential of immersive technologies (VR) to present these XAI outputs in a way that is intuitively understandable for clinicians.

Section 5.0 synthesises these findings to identify the specific research gap this project will address: the lack of an integrated, 3D-visualised expalinability pipeline for brain tumour segmentation.

section 6.0 explains the investigation's findings and explains how this initiative will fill the identified gap.

2 3D Deep Learning for Medical Brain Tumour Segmentation

2.1 Evaluation of 3D CNNs for Brain Tumour Segmentation

Over the past decade, the field of medical image analysis has experienced a significant paradigm shift, moving from **"reasoning-based"** systems that relied on manual feature engineering to **"learning-based"** Deep Learning (DL) approaches (Neri et al., 2023). In the past, radiomics was used for the automated analysis of brain tumours. To train classifiers like Support Vector Machines (SVMs), domain specialists manually retrieved quantitative features including texture, intensity, and shape. However, this method was labour-intensive, prone to observer bias, and frequently failed to capture the intricate, non-linear spatial relationships present in volumetric data (Bhati et al., 2024).

The introduction of Convolutional Neural Networks(CNNs) brought an and of manual paradigm. CNN prefer an end-to-end approach to extract features automatically by utilizing raw pixel data in contrast to classical machine learning. Initial implementation in neuro-oncology treated Magnetic Resonance imaging(MRI) as a series of 2D slices. As this approach allowed to take pre-trained networks into consideration(such as VGG or ResNet), it has significant drawback: it neglect the spatial context along the Z-axis(depth).

In neuro-oncology, a tumor is an inherently volumetric entity; it defines the anatomical boundaries with continuation across the sagittal, coronal and axial planes. When 3D volume is processed as a series of 2D slices it frequently results in **"discontinuous predictions"**, which produce uneven and anatomically impossible 3D models where lesion is found in one slice but overlooked in the adjacent one(Iftikhar et al., 2025). To overcome this problem, the field has moved towards 3D CNNs, which simultaneously conduct convolutions on the entire (X, Y, Z) volume using volumetric kernels (such as $3 * 3 * 3$). This volumetric approach allows model to learn complex spatial hierarchies and inter-slice connections thanks to its volumetric methodology, which is essential for precise segmentation of brain tumor using high scaled Dataset like BraTS.

2.2 Key 3D CNN Architectures for Brain Tumour Segmentation

To handle the Computational complexity of 3D data while conducting segmentation that is pixel-perfect, the research community has standardize the **"Encoder-Decoder"** Architecture.

2.2.1 The 3D U-Net Architecture

Building upon the success of 2D U-Net, Çiçek et al. (2016) introduce the 3D U-Net architecture, a fully convolutional network for dense volumetric images segmentation. A U-shaped Symmetric model consists of two distinct pathways:

1. **The Contracting Path(Encoder)**: This pathway significantly consider feature extraction. It consists of max-pooling layers after recursive block of 3D convolutions. As data processed through the encoder, the spatial resolution decreases while the number of feature increases.This path allows model to capture high-level conceptual data context("where is tumor present", such as differentiate tumor tissues for healthy one (Çiçek et al., 2016).
2. **The Expanding path(Decoder)**: To regenerate the segmentation map, the spatial resolution needs to be restored which has lost during encoding. This path reconstruct the feature maps back to its initial input dimensions using up-convolution(transposed convolution) layers. Also provide precise location("where is the tumor").

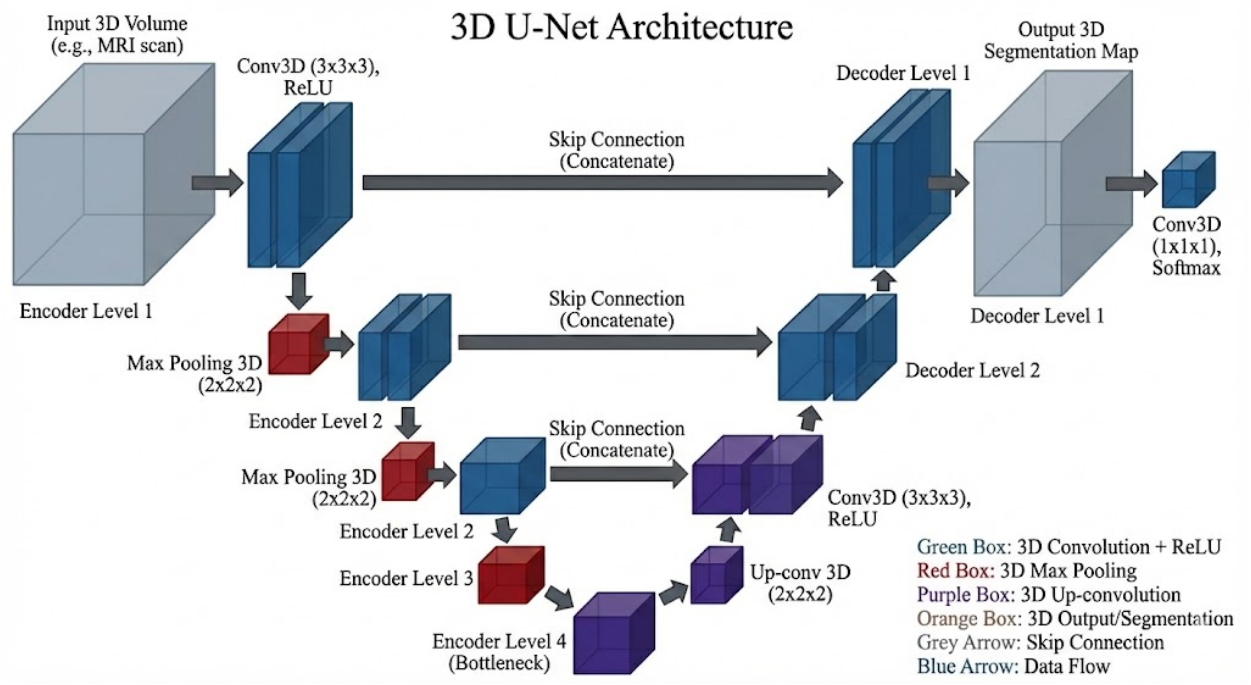

Figure 1: 3D U-Net Architecture

A critical innovation of the 3D U-Net introduce long-range skip connections. Deep neural network often face difficulty to acquire fine-grained spatial information after applying multiple pooling operations. The 3D U-Net addresses this problem concatenates high-resolution feature maps straight from the encoder to the appropriate decoder layers. According to the Çiçek et al. (2016), this mechanism allows the network to incorporate deep semantic features with shallow details,with precise boundary delineation even when it trained with sparse annotations.

2.2.2 Attention Mechanisms in 3D CNNs (3D AttUNet)

Medical images often contains irrelevant background information such as non-cosiderable tissue in the brain, while the standard U-Net treats all the pixels in a feature map equally. To overcome this issue, researchers have integrated "Attention Mechanisms" into the U-Net architecture. As Explained by Natekar et al. (2020), the Attention U-Net incorporates **"Attention Gates"** into the skip connections.

This gates utilize the coarser signal form the gating vector(the deep layer) to filter the activations from the input signal (the shallow layer) before they are merged. This enhance the activations in the region of interest(ROI) and suppresses the irrelevant background regions using mathematics equations. This "soft-attention" mechanism improves the model's sensitivity to small, irregular lesions such as glioma subregions without requiring additional guidance or complex cropping preprocessing (Natekar et al, (2020)).

2.2.3 Other Relevant Architectures: V-Net and Residual Learning

parallel to the U-Net Milletari et al.(2016) proposed the V-Net,an optimized model only for volumetric medical data. V-Net is completely distinct for other models,incorporating the **Residual connections**(short-skip connections) block. This model resembles the ResNet philosophy, add the input of a block to its output, allowing gradients to flow more smoothly during backpropagation. compared to convolutional U-Net, this reduces the vanishing gradient issue and make it easier to train much deeper 3D networks (Milletari et al., 2016).

Furthermore, the V-Net research has presented a solution to one of the most enduring challenges in the medical segmentation: Class Imbalance. In a dataset,tumor occupies less the 1% of the total volumetric pixels while rest of the area is unaffected and healthy background. Standard objective functions like Cross-Entropy refuse to get success in this scenario,results in having model achievable accuracy as 99% by predicting "background irrelevant data for every voxel. To overcome this Milletari et al. (2016) introduced the **Dice coefficient**, which optimises the model by considering the gap between the ground truth and the expected segmentation:

$$D = \frac{2\sum_i^N p_i g_i}{\sum_i^N p_i + \sum_i^N g_i}$$

where the sums run through the N voxels, of the predicted binary volume $p_i \in P$ and the ground truth binary volume $g_i \in G$. This function can improvise by rewriting it in a differentiable form. As a result, it can be optimized with gradient descent. It can be compute with respect to each predicted voxel, allowing the network to improve the overlap between the predictions and ground truth (Milletari et al., 2016).

The gradient:

$$\frac{\partial D}{\partial p_j} = 2\left[\frac{g_j}{\sum_i p_i^2 + \sum_i g_i^2} - \frac{2p_j(\sum_i p_i g_i)}{(\sum_i p_i^2 + \sum_i g_i^2)^2}\right]$$

2.3 Frameworks and Libraries

To conduct a practical implementation of the 3D architectures, specialised computational frameworks are required. while there are some general-purpose libraries like TensorFLow exists, pyTorch has dynamic computation graph (eager execution) which is dominant framework in the medical imaging research. This ability of PyTorch promotes the debugging of complex 3d tensor operations and allows for more adaptable changes to an architectures while doing research(Bhati et al., (2024)).

However, medical imaging has unique challenges that general libraries not natively address, such as diverse file formats(e.g., NifTI DICOM), managing 3D system,and perform volumetric data augmentation. To bridge this gap, the **Medical Open Network for Artificial Intelligence (MONAI)** has emerged as the standard library for medical deep learning. This library is built for domain-specific functionality on top of PyTorch, which includes:

1. **pre-built Architectures:** optimized models for medical image analysis.[3D U-Net, V-Net, and UNETR (UNet Transformers).
2. **Volumetric Transformations:** Specialized augmentations tools like as random 3D elastic deformations and intensity shifting, which are significant on small medical datasets.
3. **Sliding Window Inference:** Enables model to handle multiple 3D volumes at once. It divides huge data into patches and stitches and control the GPU memory.

MONAI guarantees consistency by abstracting the low-level tensor manipulation required for volumetric processing, allowing researchers to concentrate on high-level task such incorporating XAI modules into segmentation pipeline.

3 Explainable AI (XAI) for Explainability in Medicine

3.1 The Need for Explainability in Clinical practice

The integration of Artificial Intelligence (AI) into medical practice especially in high-stakes domains like neuro-oncology, has made it more difficult by the opacity of modern Deep learning (DL) models. Convolutional neural network has demonstrated super performance in segmentation like tasks, where their complex,non-linear decision-making remain inaccessible. This "Black Box" nature has became barrier in model adoption, often referred to as the **"trust gap"** (Wen et al., 2025).

Just model prediction is insufficient in clinical practice, it must incorporate with justification that follows medical knowledge. As highlighted by Neri et al. (2023), the ethical and legal framework governing medicine, such as "Right to Explanation" under GDPR, demand that automated decisions be transparent and contestable. clinicians must be able to identify the reason behind the model classification on a specific region as tumor present or not to ensure patient safety(Iftikhar et al., 2025).

Furthermore, XAI is not just a compliance tool but also a critical mechanism for model debugging and knowledge discovery. Researchers can detect "clever Hans" events, in which model learns misleading connection(e.g.,forecasting tumor based on specific hospital watermark rather then anatomical pathology)(Hou et al., 2024). when it comes to brain tumor segmentation,XAI refer to the validation of network focus on biologically relevant sub-regions such as the necrotic core or enhancing tumor(Natekar et al., 2020).

3.2 Taxonomy of XAI Methods

The literature categorizes XAI approaches according to their scope and timing relative to model training.

3.2.1 Post-hoc vs. Ante-hoc Interpretability

A primary differences made between Ante-hoc(or "intrinsic) and post-hoc methods.

- **Ante-hoc:** is designed to interpretable by its nature. Example, Decision Tree or Linear Regression models, where the relationships between them is explicit(Agrawal et al., 2025). By integrating interpretable prototypes blocks directly into the network architecture, recent advancement in "self-Explainable AI (S-XAI)" planning to introduce the transparency in deep learning (Hou et al., 2024).
- **post-hoc** methods are designed for explainability after the model has been trained. they treat a model as a "black box" and try to improve and approximate its behaviour based on the given inputs and outputs. This method prefers to be standard for analyzing medical imaging for model like 3D U-Net architecture to perform segmentation(Bhati et al., 2024),

3.2.2 Local vs. Global Explanation

Further XAI methods are divided by their scope:

- **Global Explanations**: refers to methods that explain the overall logic of the model across the entire dataset, such as feature importance in model.
- **Local Explanations**: refers to methods that explain the decision-making process of the model for specific inputs, such as reason behind by classified voxel region as tumor. Local explanation is important for clinical decision making as they offer the patient-specific rationale to diagnosis and therapy planning (Iftikhar et al., 2025).

3.3 Review of Voxel-Based XAI Techniques

For 3D medical imaging, the most common XAI techniques are those that can generate the saliency maps, spatial representations of "feature importance". The three most prominent approaches are Grad-CAM, LIME and SHAP.

3.3.1 Gradient-Based Methods: Grad-CAM

Gradient-weighted Class Activation Mapping(Grad-CAM) has recognize as the "Golden Standard" for the visualizing CNN decision in the medical imaging (Bhati et al., 2024). Grad-CAM is model-agnostic and can be applied to any CNN models with retraining compare to the earlier methods that required architectural changes.

Mechanism: Grad-CAM utilize the gradient of the target concept(e.g., tumor) to produce a saliency map, which highlights the regions of the input image that contribute most to the prediction of the target class. It computes a weighted sum of the feature maps, where the weight represent the importance of each feature that channel to the prediction. To eliminate the negative contributions,A Rectified Linear Unit (ReLU) is then applied highlighting more specific regions that support positively the class of interest (Selvaraju et al., 2017).

Application in 3D: Natekar et al.(2020) successfully applied Grad-CAM to a 3D U-Net for brain tumour segmentation as it was originally designed for 2D images. There research showed that 3D Grad-CAM can localize tumor based on the aggregated gradients across volumetric feature maps. However, it has critical limitation which is resolution trade-off:because Grad-CAM uses feature maps from deep layers. As a result, the resulting heatmaps are often coarse and may bleed into surrounding tissue, reducing the voxel-level precision of the segmentation mask itself (Natekar et al., 2020).

3.3.2 Perturbation-Based Methods: LIME

Local Interpretable Model-agnostic Explanations (LIME) has different approach. LIME treats the model as function only, without using internal Gradients.

Mechanism: LIME modifies the model's predictions and observe the change by masking out random superpixels(segments). It then fits a simple, interpretable model to this disrupted samples to predict model's behavior locally around specific image (Ribeiro et al., 2016).

Evaluation in 3D: while LIME is highly recommended for 2D image segmentation, its performance in 3D is computationally expensive. According to the Agrawal et al., (2025) LIME require thousand of forward passes to generate stable explanation. For data like 3D images which has millions of voxels, generating valid superpixels and running sufficient distribution to achieve the significant statistics is often too slow for real-time clinical workflows.Additionally, to define the "superpixels" in a 3D brain volume is quite difficult, frequently leads to unstable explanations that vary between runs.

3.3.3 Game-Theoretic Methods: SHAP

SHAP (SHapley Additive exPlanations) provide a theoretically sound substitute based on cooperative game theory. Each feature(voxel) is given a "importance value" that represents its marginal contributions to the prediction, averaged over all possible combinations of features(Ribeiro et al., 2016).

Strength and Weaknesses: SHAP provide the most mathematically consistent explanations, by conforming the summation of each feature attributes that is equal to the model's output. However, it has extreme computational costs in high-dimensional spaces. "DeepSHAP," an approximation method, has been used to accelerate this process, but research suggest that SHAP values can mislead the Explanation in deep networks, as they may violate axioms when features are highly correlated, which is often common in statically coherent MRI data (Miró-Nicolau et al., 2024).

3.4 Comparative Synthesis and the Fidelity Gap

All three methods have their strengths and weaknesses, but Grad-CAM is still the best performing method for 3D medical image segmentation because of its computational efficiency (only one backward pass is needed) and its capacity to make use use of the spatial information present in CNN feature maps (Natekar et al., 2020).

However, a critical "fidelity gap" exist, as existing measures frequently show significant discrepancy between the predicted and ground truth segmentation (Miró-Nicolau et al., 2024). This discrepancy

is exacerbated by the standard practice of analyzing 3D attention maps as static 2D slices, this is the method that hides volumetric coherence and increase cognitive load for clinicians. Therefore, closing the "trust gap" calls for a paradigm change towards immersive Virtual Reality, which provides the high-dimensional, intuitive visualization needed to properly understand these deep volumetric insights, rather then simply algorithmic improvements.

4 Immersive Visualization in Medical Imaging

4.1 Traditional 2D vs. Immersive 3D Visualization

The interpretation of volumetric medical data has historically been analyse on 2D screens, which causes a lot of cognitive stress. According to Khedir et al. (2025) traditional slice-by-slice navigation requires to mentally reconstruct in volumetric architecture through clinicians. This process requires time as well as prone to spatial errors when you assessing the depth of the tumors relative to prominent brain regions. Even though 3D Convolutional Neural Network (CNNs) can process this volumetric data,the representation of their outputs remains largely stuck in 2D.

Emerging research suggest that immersive visualization such as Virtual Reality(VR) and Augmented Reality(AR) can bridge this gap. VR provide **stereoscopic depth perception** and **6-Degrees-of-Freedom(6DoF)** interaction which is way beyond then standard monitors. This allow user to view the "saliency maps" not as flat overlay which is generated by the XAI but gives volumetric clouds suspended in 3D space. The NeuroXAI framework has immersive environments significantly improves the clinician's ability to localize pathology and understand complex neural connectivity compared to standard desktop viewers(Zeineldin et al., 2022).

4.2 VR as a Tool for Surgical Planning and Medical Education

The application of VR is extends way beyond from passive viewing to active surgical planning. Defining tumor boundaries in neuro-oncology is critical. Recent advancement in Augmented Reality (AR) enables surgeons simulate trajectories and visualize "risk maps" created by deep learning models before entering to the operation theatre(Zeineldin, 2023; Khedir et al., 2025) . For situation like complex glioma, where 2D scans may not represent the tumor with key blood arteries, this tool is crucial resection corridor.

Additionally, the idea of interactive AI is essential to current medical systems. Static heatmaps (like standard Grad-CAM) offers **"take it or leave it"** explanation. VR environments, on the other hand make the **human-in-the-loop(HITL)** interaction easier. immersive interfaces enable clinicians to interrogate the model e.g. by digitally "erasing" a portion of the input volume to identify the prediction changes (Orsmaa et al., 2025). The interactivity change XAI from a static report into dynamic conversation between clinician and the AI, allowing the correction of model errors in real time(Orsmaa et al., 2025).

4.3 Technologies for VR Visualization: 3D Slicer and SlicerVR

To implement this immersive systems requires strong technological pipeline. Standard medical formats (DICOM/NIfTI) are not supported by native game engines like Unity. However, platforms like 3D slicer have fill the bridge. SlicerVR extension allows volume rendering of medical scans in VR headsets without any data conversation loss. Despite this, VR still continues to be a major difficulty in terms of integrating XAI. Most of the pipelines concentrate on 2D slice-based visualization,which fails to convey the volumetric feature importance and limits the clinical utility of XAI in more in depth scenarios (Zeineldin et al., 2022). There is a need for distinct workflows that can consume a 3D CNN's attention weights and render them as interactive volumetric objects, as proposed in frameworks like DeepIGN (Zeineldin, 2023).

5 Synthesis and Research Gap

5.1 The Disconnect Between Model Accuracy and Clinical Utility

After addressing and reviewing three most significant chapters of this review, A distinct contrast has been established. The models like 3D U-Net and its variants has achieved a tremendous success in segmenting brain tumor (Çiçek et al., 2016; Milletari et al., 2016). On the other side the opacity of the "Black Box" remains a critical barrier to the clinical trust (Neri et al., 2023). Although methods like Grad-CAM have successfully illuminated the internal logic of these networks (Natekar et al., 2020), current implementation suffers from three major limitations that this project aims to address.

5.2 Identified Gaps in Current XAI Frameworks

Having global achievements in 3D XAI based on research such as NeuroXAI, there is still something which is called **"interactivity gap"** exists; current system static visualization rather than an active contribution of a user (clinician)(Zeineldin et al., 2022). This passivity leads to a serious methodological error that confuses feature relevance with model confidence. As mentioned in the DeepIGN study (Zeineldin, (2023)), a network may exhibit high focal attention on a specific region while displaying high epistemic uncertainty. This could give confident result but incorrect false positives. As a result, existing 3D XAI methods do not provide the required mechanisms for **Interactive Correction or Uncertainty Quantification**, leave the functional gap where user cannot distinguish between a confident prediction and a statistical guess, or to change the model's focus (Abyasa and Rahmania, 2025).

Furthermore, the clinical utility of these volumetric insights is severely constrained by the "cognitive friction" of traditional visualization mediums. Khedir-Amara et al (2025) mentioned that analyzing 3D volumetric pathology via slice-by-slice 2D navigation increases significant burden on clinicians to mentally reconstruct spatial relationships from the disconnected images. This 3D nature of the data and the 2D nature of the display limits the interpretability of the feature maps, as crucial depth signals lost in translation. Therefore, a clear research exists for a need of unique framework that not only generates, uncertainty-aware explanation but also represents that in an immersive virtual environment, to clarify whether or not the engagement significantly lowers the cognitive load as compare to conventional medical workflows.

6 Conclusion

This literature review establishes while 3D Deep learning architectures, particularly the 3D U-Net, have achieved superior success in volumetric tumor segmentation, the deployment is critically hindered by the "Black Box" paradox (Neri et al., 2023). The clinical workflows still remain detach even if the Explainable methods like Grad-CAM provide mathematical solution for transparency. The reliance on static 2D slice-by-slice visualization fails to convey the complex spatial resolution of 3D pathologies, is hindered by a significant cognitive load that limits the clinical trust (Natekar et al., 2020).

Consequently, this research focusing on bridging the "trust gap" requires the paradigm shift form passive observation to the interactive environment. By developing a virtual reality environment that incorporate uncertainty quantification with real time feedback, this project aims to "Illuminate the Black Box" effectively. This innovative method will ensure that this High-performance AI is not just accurate but also understandable and useful for medical practice.

7 References

Abyasa, J. and Rahmania, R. (2025) "AXONS-3: An XAI-augmented approach for advancing trust and transparency in 3D brain tumor segmentation," in, pp. 64–71. Available at: https://doi.org/10.1109/CogSIMA64436.2025.11079472.

Agrawal, R. *et al.* (2025) "Fostering trust and interpretability: Integrating explainable AI (XAI) with machine learning for enhanced disease prediction and decision transparency," *Diagnostic Pathology*, 20. Available at: https://doi.org/10.1186/s13000-025-01686-3.

Bhati, D., Neha, F. and Amiruzzaman, M. (2024) "A survey on explainable artificial intelligence (XAI) techniques for visualizing deep learning models in medical imaging," *Journal of Imaging*, 10. Available at: https://doi.org/10.3390/jimaging10100239.

Çiçek, Ö. *et al.* (2016) "3D u-net: Learning dense volumetric segmentation from sparse annotation," in S. Ourselin et al. (eds.) *Medical image computing and computer-assisted intervention – MICCAI 2016*. Cham: Springer International Publishing, pp. 424–432.

Hou, J. *et al.* (2024) "Self-eXplainable AI for medical image analysis: A survey and new outlooks." Available at: https://arxiv.org/abs/2410.02331.

Iftikhar, S. *et al.* (2025) "Explainable CNN for brain tumor detection and classification through XAI based key features identification," *Brain Informatics*, 12. Available at: https://doi.org/10.1186/s40708-025-00257-y.

Khedir, M. *et al.* (2025) "BrainAR: Automated brain tumor diagnosis with deep learning and 3D augmented reality visualization," *IEEE Access*, 13, pp. 128639–128653. Available at: https://doi.org/10.1109/ACCESS.2025.3590291.

Menze, B.H. *et al.* (2015) "The multimodal brain tumor image segmentation benchmark (BRATS)," *IEEE Transactions on Medical Imaging*, 34(10), pp. 1993–2024. Available at: https://doi.org/10.1109/TMI.2014.2377694.

Milletari, F., Navab, N. and Ahmadi, S.-A. (2016) "V-net: Fully convolutional neural networks for volumetric medical image segmentation," in *2016 fourth international conference on 3D vision (3DV)*, pp. 565–571. Available at: https://doi.org/10.1109/3DV.2016.79.

Miró-Nicolau, M., Jaume-i-Capó, A. and Moyà-Alcover, G. (2025) "A comprehensive study on fidelity metrics for XAI," *Information Processing & Management*, 62(1), p. 103900. Available at: https://doi.org/https://doi.org/10.1016/j.ipm.2024.103900.

Natekar, P., Kori, A. and Krishnamurthi, G. (2020) "Demystifying brain tumour segmentation networks: Interpretability and uncertainty analysis." Available at: https://arxiv.org/abs/1909.01498.

Neri, E. *et al.* (2023) "Explainable AI in radiology: A white paper of the italian society of medical and interventional radiology," *La Radiologia medica*, 128. Available at: https://doi.org/10.1007/s11547-023-01634-5.

Orsmaa, L. *et al.* (2025) "Interactive AI annotation of medical images in a virtual reality environment," *International Journal of Computer Assisted Radiology and Surgery* [Preprint]. Available at: https://doi.org/10.1007/s11548-025-03497-9.

Ribeiro, M.T., Singh, S. and Guestrin, C. (2016) ""Why should i trust you?": Explaining the predictions of any classifier." Available at: https://arxiv.org/abs/1602.04938.

Selvaraju, R.R. *et al.* (2017) "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *2017 IEEE international conference on computer vision (ICCV)*, pp. 618–626. Available at: https://doi.org/10.1109/ICCV.2017.74.

Wen, B. *et al.* (2025) "Towards a transparent and interpretable AI model for medical image classifications," *Cognitive Neurodynamics*, 19(1). Available at: https://doi.org/10.1007/s11571-025-10343-w.

Zeineldin, R. (2023) *Deep multimodality image-guided system for assisting neurosurgery.* PhD thesis. Informatik. Available at: https://doi.org/10.5445/IR/1000155782.

Zeineldin, R.A. *et al.* (2022) "Explainability of deep neural networks for MRI analysis of brain tumors," *International Journal of Computer Assisted Radiology and Surgery*, pp. 1–11.