Advances in Clinical Medicine 临床医学进展, 2026, 16(3), 546-553
Published Online March 2026 in Hans. https://www.hanspub.org/journal/acm
https://doi.org/10.12677/acm.2026.163821

# 可解释人工智能在脑肿瘤的应用概述

林国鉴[1]，李珍珠[2]，刘志鹏[2]，油亚倩[2]，刘书勇[1]，王波定[2*]

[1]绍兴文理学院医学院，浙江 绍兴
[2]宁波市第二医院神经外科，浙江 宁波

收稿日期：2026年1月27日；录用日期：2026年2月22日；发布日期：2026年3月3日

## 摘 要

**脑肿瘤作为一种侵袭性强、复发率高的疾病，对患者的生存率和生活质量有着显著影响。现代医学影像技术如MRI和CT为脑肿瘤的检测和诊断提供了宝贵的支持，而近年来兴起的人工智能(AI)和深度学习技术更是推动了脑肿瘤分析的自动化与智能化发展。然而，深度学习模型的“黑箱”性质限制了它们在实际医疗环境中的广泛应用，临床医生难以理解或信任这些缺乏解释性的模型。可解释人工智能(XAI)应运而生，旨在提升深度学习模型的透明性和可理解性，从而增强其在脑肿瘤检测、分类和预后预测中的应用效果。本文综述了当前XAI在脑肿瘤诊断中的应用进展，探讨了不同XAI方法的优缺点及其在临床场景中的适用性，总结了XAI在医学影像分析中的挑战与未来发展方向。**

## 关键词

**脑肿瘤，可解释人工智能(XAI)，深度学习**

# A Review of the Applications of Explainable Artificial Intelligence in Brain Tumors

**Guojian Lin[1], Zhenzhu Li[2], Zhipeng Liu[2], Yaqian You[2], Shuyong Liu[1], Boding Wang[2*]**

[1]School of Medicine, Shaoxing University, Shaoxing Zhejiang
[2]Department of Neurosurgery, Ningbo NO.2 Hospital, Ningbo Zhejiang

Received: January 27, 2026; accepted: February 22, 2026; published: March 3, 2026

## Abstract

**Brain tumors are highly aggressive and have high recurrence rates, significantly affecting patients' survival and quality of life. Modern medical imaging technologies, such as magnetic resonance imaging**

*通讯作者。

文章引用: 林国鉴, 李珍珠, 刘志鹏, 油亚倩, 刘书勇, 王波定. 可解释人工智能在脑肿瘤的应用概述[J]. 临床医学进展, 2026, 16(3): 546-553. DOI: 10.12677/acm.2026.163821

**(MRI) and computed tomography (CT), provide valuable support for the detection and diagnosis of brain tumors. In recent years, the emergence of artificial intelligence (AI) and deep learning has further promoted the automation and intelligent analysis of brain tumors. However, the "black-box" nature of deep learning models limits their widespread adoption in real clinical environments, as clinicians find it difficult to understand or trust models that lack interpretability. Explainable artificial intelligence (XAI) has therefore emerged to enhance the transparency and interpretability of deep learning models, thereby improving their applicability in brain tumor detection, classification, and prognosis prediction. This article reviews the current progress of XAI in brain tumor diagnosis, discusses the advantages and limitations of different XAI methods and their suitability for clinical scenarios, and summarizes the challenges and future directions of XAI in medical imaging analysis.**

## Keywords

**Brain Tumor, Explainable Artificial Intelligence, Deep Learning**



## 1. 引言

脑肿瘤分为原发性脑肿瘤和转移性脑肿瘤两大类。由于其对神经功能的直接损害，脑肿瘤的早期诊断和准确分类对制定治疗方案和提高患者生存率至关重要[1]。然而，脑肿瘤的异质性较强，不同类型的肿瘤在影像特征上差异明显，导致传统的影像诊断面临诸多挑战。为克服这些挑战，现代医学影像分析正朝着人工智能驱动的方向发展，尤其是深度学习在自动化分析和图像处理方面的巨大潜力，给脑肿瘤的检测和诊断带来了前所未有的机遇[2]。

### 1.1. 脑肿瘤的诊断现状与挑战

当前，医学影像是脑肿瘤检测和诊断的主要手段。典型的影像学检查包括磁共振成像(MRI)、计算机断层扫描(CT)等，这些技术能够提供脑部结构和功能信息，辅助医生判断肿瘤的位置、大小和性质[3]。然而，传统影像分析高度依赖于放射科医生的专业知识和经验，这使得诊断结果在一定程度上受到主观因素的影响，尤其在脑肿瘤的类型和恶性程度存在不确定性时。此外，放射科医生通常需要对大量图像逐一分析，这种高强度的工作导致误诊风险增加和诊断效率降[4]。

近年来，人工智能技术，特别是深度学习，在图像识别、特征提取和模式识别方面展现了出色的性能，使得脑肿瘤的自动化检测和分类成为可能。基于深度学习的算法能够通过对大量标记图像数据的学习，提取肿瘤区域的特征，从而在检测和分类准确性上逐步接近甚至超越人类专家的水平。然而，传统深度学习模型的“黑箱”性质，即难以解释模型做出决策的具体过程，阻碍了其在临床中的广泛应用[5]。临床环境要求算法结果具有可解释性，以便医生能理解和验证模型的判断依据。可解释性成为了人工智能在医疗领域应用的关键需求[6]。

### 1.2. 可解释人工智能的兴起

为解决深度学习模型的“黑箱”问题，可解释人工智能(XAI)作为一个新兴领域得到了快速发展[7]。XAI 的核心思想是使人工智能模型的决策过程变得可理解和透明，允许使用者了解模型如何对输入数据进行处理、分析并生成输出结果。通过 XAI 技术，医学影像分析模型可以将关注的特征以可视化的形式

呈现给医生，帮助他们理解模型的判断依据，从而增强对模型结果的信任[5] [8]。

可解释人工智能的应用不仅提升了模型的透明度，还能提高其在医学影像分析中的应用价值。对于脑肿瘤等复杂病症，XAI 可以通过可视化热力图等方式展示模型关注的影像区域，帮助医生确认模型的重点区域是否与病灶一致[9] [10]。此外，通过 XAI 方法能够揭示不同特征对诊断结果的贡献度，为个性化医疗和精确诊断提供支持。因此，可解释人工智能在提升医疗 AI 的可信度和实用性方面具有重要的意义。

### 1.3. XAI 在脑肿瘤诊断中的重要性

在脑肿瘤的检测和分类中，XAI 的作用尤为突出。不同类型的脑肿瘤在影像特征上的表现差异显著，例如胶质瘤和脑膜瘤在形态、边缘等方面的表现不同，传统的深度学习模型虽然能对这些特征进行分类，但其具体的特征判断依据却难以理解[11]。可解释 AI 方法如 Grad-CAM、LIME 和 SHAP 等，通过对模型的注意力机制或特征贡献进行解释，使得医学影像分析模型的决策过程逐渐透明化。这些解释技术可以帮助医生验证模型的判断是否符合实际病理特征，从而提升诊断的准确性和可靠性。

当前，越来越多的研究聚焦于 XAI 在医学影像分析中的应用。可解释 AI 技术如 Grad-CAM 和 LIME 常用于高分辨率的 MRI 图像，通过生成热点图来展示模型“关注”的区域。对脑肿瘤图像来说，这些热点图不仅可以帮助医生直观了解模型的决策依据，还可以指出模型在肿瘤区域的识别和分类方面是否出现偏差[12]。随着这些技术的发展，XAI 在脑肿瘤诊断中的应用前景十分广阔，越来越多的研究致力于开发更加精准、高效的解释方法，以提升医学人工智能的可用性和临床价值。

## 2. 可解释人工智能的概述

近年来，随着深度学习和机器学习模型的广泛应用，可解释人工智能(XAI)逐渐成为人工智能研究中的一个重要分支。在医学领域，XAI 的出现不仅促进了 AI 技术在医学影像分析中的应用，也在很大程度上提升了模型的可信度。不同于传统的“黑箱”模型，XAI 的核心在于使 AI 模型的决策过程透明化，使医生能够理解并验证模型的判断依据。在医学影像分析中，XAI 为肿瘤检测、分级、分类等应用场景带来了广泛的前景。以下将对 XAI 的基本概念、主流方法及其在医学影像中的应用进行详细介绍。

### 2.1. XAI 的主要分类

XAI 技术根据其解释方式和应用场景的不同，可以分为几类主要方法：可视化解释、特征重要性分析、局部解释模型等。以下是几种常用的 XAI 方法。

#### 2.1.1. 可视化解释

可视化解释是 XAI 应用最广泛的方式之一，尤其适用于图像数据。它主要通过生成关注热力图，展示模型在分类或预测过程中关注的图像区域。可视化解释的代表性方法包括 Grad-CAM (Gradient-weighted Class Activation Mapping)和 LIME (Local Interpretable Model-agnostic Explanations)。

Grad-CAM：Grad-CAM 是一种基于卷积神经网络(CNN)的可视化技术，通过计算图像中每个像素对预测结果的梯度，生成一个关注区域的热力图[13]。Grad-CAM 尤其适用于医学影像分析，因为它可以直观地展示模型关注的重点区域。研究表明，Grad-CAM 在 MRI 图像的脑肿瘤检测中表现出色，通过热力图帮助医生识别肿瘤的具体位置[14]。

LIME：LIME 通过对模型的预测进行局部线性近似，从而生成特定区域的解释。它通过对原始图像的局部区域进行扰动，观察模型预测结果的变化，从而推断出不同区域对预测结果的贡献。LIME 适用于任意类型的机器学习模型，尤其在多模态影像数据的处理上表现出色[5] [15]。

#### 2.1.2. 特征重要性分析

特征重要性分析方法主要用于揭示不同特征对模型预测结果的影响。在脑肿瘤诊断中，不同的影像特征(如形状、纹理、强度等)对肿瘤的检测和分类具有不同的贡献，因此通过特征重要性分析可以帮助医生更好地理解模型的决策过程[16]。代表性方法包括 SHAP (SHapley Additive exPlanations)和基于神经网络权重的特征重要性分析。

SHAP：SHAP 基于博弈论中的 Shapley 值，通过对每个特征的贡献进行量化，揭示其在预测中的重要性。在医学影像分析中，SHAP 可以帮助医生了解模型是如何基于不同的影像特征做出诊断决策的。对于脑肿瘤诊断，SHAP 可以展示每个影像特征(例如纹理、密度)的重要性，使得模型在解释肿瘤类型和级别上更具透明性[17]。

#### 2.1.3. 局部解释模型

局部解释模型通过在输入数据的局部区域进行分析，生成局部解释。相比于整体解释方法，局部解释更加灵活，可以在不同场景中提供更为细致的解释。例如，在多模态 MRI 图像中，不同的影像层可以通过局部解释模型分别分析，帮助医生理解模型如何整合多层信息[18]。

对比敏感度映射：此类方法通过对图像的局部区域进行微小的扰动或掩盖，观察模型预测结果的变化，推断出该区域对结果的重要性。对比敏感度映射适合于脑肿瘤的图像分析，尤其在确定肿瘤的边界和关键区域方面，能够提供有用的信息[13]。

## 3. 可解释深度学习的发展

随着深度学习在医学影像分析领域的迅速普及，研究人员不断探索将深度学习模型与可解释 AI 方法相结合，以提升模型在医学应用中的可靠性。深度学习模型，特别是卷积神经网络(CNN)和基于视觉转换器(Vision Transformer, ViT)的网络结构，广泛应用于脑肿瘤的检测和分类任务[20]。深度学习通过自动提取高维特征，能够有效捕捉医学影像中的病灶区域，提升了模型在脑肿瘤分类、检测和预后预测中的表现。然而，深度学习模型的复杂性导致其内部决策过程难以直观理解，因此需要借助 XAI 方法为其赋予解释性[19]。以下内容将深入分析几种典型深度学习模型及其在医学影像分析中的可解释性发展。

### 3.1. 深度学习模型的结构和特点

卷积神经网络(CNN)是最常用于图像分析的深度学习模型之一，其结构能够捕获图像中的空间特征。CNN 主要由卷积层、池化层和全连接层组成，通过逐层卷积提取图像特征，并在后续层级中逐渐聚合特征信息，从而完成对图像的分类或分割。CNN 的卷积运算能够有效地减少计算量，使其在医学图像处理任务中非常高效[10]。对于脑肿瘤的检测，CNN 能够在 MRI 或 CT 图像中识别肿瘤区域，帮助医生进行病灶定位。然而，CNN 模型的解释性较差，其卷积层的决策机制在临床环境中难以理解。

视觉转换器是一种基于自注意力机制的深度学习模型，近年来在图像处理任务中取得了显著进展。与 CNN 不同，ViT 不依赖于卷积运算，而是通过自注意力机制捕捉图像的全局特征，这使得其在处理大型图像数据集时具有更强的特征提取能力[14]。ViT 模型通过对输入图像进行分块处理，从每一块中提取信息，并通过自注意力机制将各块信息进行加权汇总，从而获取全局特征。ViT 在脑肿瘤的分类任务中具有很高的准确率，尤其在异质性较高的肿瘤类型分类中表现突出。然而，由于自注意力机制的复杂性，ViT 模型的内部决策过程难以直观解释，因此通常结合 Grad-CAM 等方法进行可视化，以生成热力图显示模型关注的区域。

混合深度学习模型是近年来发展起来的一种新型结构，将 CNN 和 ViT 等不同网络架构的优势进行融合。这类模型通常采用 CNN 进行初步特征提取，随后通过自注意力机制进行全局信息聚合，以实现更

精准的图像分类和检测。在脑肿瘤诊断中，混合模型被用于提高模型的鲁棒性和特征捕捉能力。例如，研究人员通过混合 ViT 与卷积层，实现了对脑肿瘤的多层次特征提取，有效提升了检测和分类的准确性。对于这种复杂的混合模型，研究人员借助 Grad-CAM 和 SHAP 方法对其进行可解释性增强，从而展示模型在多层次特征提取中的关注区域和重要特征[16]。

### 3.2. 深度学习模型的常用解释方法

Grad-CAM 是一种基于梯度的可视化方法，其核心思想是通过计算图像中每个像素对模型预测结果的梯度，生成关注区域的热力图。在脑肿瘤检测中，Grad-CAM 可以直观地显示模型“关注”的图像区域，帮助医生确认模型是否识别出关键病灶[9]。研究表明，Grad-CAM 在 MRI 图像的肿瘤检测中表现优异，例如在检测胶质瘤等高度侵袭性肿瘤时，Grad-CAM 生成的热力图能够突出显示肿瘤边界区域，使医生更直观地理解模型的决策过程[21]。

LIME 是一种局部解释方法，通过在原始图像的局部区域进行扰动，观察模型预测结果的变化，从而推断出不同区域对预测结果的贡献。LIME 可以针对任何深度学习模型进行解释，尤其适用于分析 MRI 图像的局部特征。对于脑肿瘤的分类任务，LIME 能够生成特定区域的解释，帮助医生识别重要的局部特征，例如肿瘤内部的纹理和密度差异[22]。LIME 在临床应用中表现出良好的

SHAP 是一种基于博弈论的特征贡献度分析方法，通过计算每个特征对模型预测的贡献值，为每个输入特征分配一个 Shapley 值，从而量化其对模型预测结果的重要性[11] [23]。SHAP 的优势在于其解释结果具有一致性和可重复性。对于脑肿瘤的预后预测，SHAP 方法可以展示不同影像特征对生存期预测的贡献，为医生在制定治疗方案时提供重要参考[24]。研究显示，使用 SHAP 对深度学习模型进行解释，能够揭示不同影像特征在预后预测中的相对重要性，为个性化医疗提供了支持。

## 4. 可解释 AI 在脑肿瘤的应用

### 4.1. 脑肿瘤检测中的可解释性 AI 应用

脑肿瘤的早期检测对患者的预后有着至关重要的影响。然而，由于脑部结构的复杂性和肿瘤特征的多样性，传统的影像学方法在识别早期病灶时可能存在局限性。基于深度学习的 AI 模型能够通过学习大量 MRI 图像中的肿瘤特征实现自动检测，但其“黑箱”性质限制了在临床中的应用。XAI 技术则通过可视化解释，使得 AI 模型的检测过程更加直观，便于医生理解和判断。

在脑肿瘤检测中，有些早期病灶表现为局部细微的特征，如影像中轻微的形态或密度变化。LIME 方法通过对局部图像区域进行扰动，观察模型输出的变化，从而生成局部解释。在脑肿瘤的早期检测中，LIME 可以帮助识别影像中的微小异常，揭示肿瘤的潜在区域，特别适用于胶质瘤和脑膜瘤等肿瘤类型的检测。医生通过 LIME 生成的解释图像，可以更精细地观察到模型“认为”存在异常的区域，这有助于发现早期的微小病灶[25]。

### 4.2. 脑肿瘤分类中的可解释性 AI 应用

脑肿瘤的准确分类对于制定治疗方案至关重要。不同类型的脑肿瘤在治疗方式和预后上差异较大，例如脑膜瘤和胶质瘤的治疗方案可能完全不同，因此准确的分类诊断能够显著影响患者的治疗效果。深度学习模型通过对 MRI 图像特征的学习能够实现肿瘤的自动分类，而 XAI 技术则使得这种分类更加透明和可靠。

SHAP 方法通过计算每个特征对分类结果的贡献度，帮助医生理解模型在不同类型肿瘤的分类依据。在脑肿瘤的分类任务中，研究人员通过 SHAP 方法分析不同影像特征对分类结果的影响，如肿瘤的密度、

纹理和形态等。通过 SHAP 图像，医生可以清楚地看到每个特征对模型最终分类的贡献度，从而验证模型的分类结果是否具有临床意义。例如，对于恶性胶质瘤和脑膜瘤的区分，SHAP 方法可以帮助医生理解模型如何基于不同的影像特征判断肿瘤的类型[26]。

### 4.3. 脑肿瘤预后预测中的可解释性 AI 应用

脑肿瘤的预后预测主要关注患者的生存时间、复发风险等因素，这对制定个性化治疗方案有重要指导意义。机器学习模型通过学习患者的临床特征、影像特征等信息，可以生成个性化的预后预测结果。XAI 技术在预后预测中同样发挥着重要作用，通过解释不同特征对预后预测的影响，为医生提供决策依据。

SHAP 方法在脑肿瘤的生存预测中被广泛应用，通过计算每个特征对生存预测的贡献度，揭示不同特征的相对重要性。例如，在胶质瘤患者的生存预测中，研究人员通过 SHAP 分析 MRI 图像的不同特征(如密度、纹理等)对生存期预测的影响。通过 SHAP 解释，医生可以清晰地看到哪些特征对生存预测贡献最大，从而帮助医生进行个性化治疗决策。例如，MRI 图像中肿瘤边缘的形态特征可能与患者的预后密切相关，而这种解释可以为医生提供更多信息。

## 5. XAI 评估指标

### 5.1. 定性评估

定性评估通常依赖临床医生或影像学专家的主观判断，通过对可解释结果(如显著性热图、注意力图等)进行人工评分，以评估其是否与医学先验知识和临床经验相一致。常见方式包括对解释结果在病灶定位准确性、边界一致性以及临床合理性等方面进行打分或等级评价[27]。该方法能够直接反映 XAI 在临床应用中的可接受程度，但受评估者经验和主观因素影响较大，重复性和客观性相对有限[28]。

### 5.2. 定量评估

定量评估则通过客观指标对解释结果进行数值化衡量，是当前研究的热点方向。其中，Pointing Game 常用于评估解释方法在病灶定位方面的准确性，其核心思想是判断模型最显著响应点是否落在真实病灶区域内；Insertion/Deletion 曲线则通过逐步插入或删除重要特征，分析模型预测性能随特征变化的趋势，从而衡量解释结果与模型决策之间的一致性[29]。此外，还有基于区域重叠度、预测置信度变化等指标的评估方法，为 XAI 的客观比较提供了量化依据[30]。

## 6. 结论

脑肿瘤的早期检测、精准分类与可靠的预后预测对患者生存率具有重要意义。尽管深度学习模型在医学影像分析中取得了显著进展，但其“黑箱”特性在很大程度上限制了临床可接受性。可解释人工智能(XAI)的引入为解决这一问题提供了有效路径。Grad-CAM、LIME 和 SHAP 等方法通过可视化解释、局部扰动分析和特征贡献度量化，使模型决策过程更加透明，有助于医生理解和验证模型判断依据。在脑肿瘤检测、分类及生存预测任务中，XAI 显著提升了模型的可信度与临床实用性。结合 CNN、ViT 及其混合架构的可解释深度学习框架，为构建高性能、高可靠性的智能诊断系统奠定了基础。未来需进一步提升解释稳定性、因果性与临床可读性，以推动 XAI 在脑肿瘤精准医疗中的广泛应用。

## 基金项目

宁波市重大任务科技攻关项目(2022Z126)。

山东省科技型中小企业创新能力提升工程项目(2023TSGC0921)。

## 参考文献


[1] Aldape, K., Brindle, K.M., Chesler, L., Chopra, R., Gajjar, A., Gilbert, M.R., *et al*. (2019) Challenges to Curing Primary Brain Tumours. *Nature Reviews Clinical Oncology*, **16**, 509-520. https://doi.org/10.1038/s41571-019-0177-5

[2] Hu, L.S., Hawkins-Daarud, A., Wang, L., Li, J. and Swanson, K.R. (2020) Imaging of Intratumoral Heterogeneity in High-Grade Glioma. *Cancer Letters*, **477**, 97-106. https://doi.org/10.1016/j.canlet.2020.02.025

[3] Dorfner, F.J., Patel, J.B., Kalpathy-Cramer, J., Gerstner, E.R. and Bridge, C.P. (2025) A Review of Deep Learning for Brain Tumor Analysis in MRI. *NPJ Precision Oncology*, **9**, Article No. 2. https://doi.org/10.1038/s41698-024-00789-2

[4] Ottoni, M., Kasperczuk, A. and Tavora, L.M.N. (2025) Machine Learning in MRI Brain Imaging: A Review of Methods, Challenges, and Future Directions. *Diagnostics*, **15**, Article No. 2692. https://doi.org/10.3390/diagnostics15212692

[5] Abraham, L.A., Palanisamy, G., Veerapu, G. and Nisha, J.S. (2025) Exploring the Potential of Explainable AI in Brain Tumor Detection and Classification: A Systematic Review. *Artificial Intelligence Review*, **59**, Article No. 14. https://doi.org/10.1007/s10462-025-11410-8

[6] 陈园琼, 朱承璋. 医学影像处理的深度学习可解释性研究进展[J]. 浙江大学学报(理学版), 2021, 48(1): 18-29.

[7] Erukude, S.T., Chaitanya Marella, V. and Veluru, S.R. (2025) Explainable Deep Learning in Medical Imaging: Brain Tumor and Pneumonia Detection. 2025 *4th International Conference on Innovative Mechanisms for Industry Applications* (*ICIMIA*), Tirupur, 3-5 September 2025, 906-911. https://doi.org/10.1109/icimia67127.2025.11200629

[8] 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**, Article No. 239. https://doi.org/10.3390/jimaging10100239

[9] Akgündoğdu, A. and Çelikbaş, Ş. (2025) Explainable Deep Learning Framework for Brain Tumor Detection: Integrating LIME, Grad-CAM, and SHAP for Enhanced Accuracy. *Medical Engineering & Physics*, **144**, Article ID: 104405. https://doi.org/10.1016/j.medengphy.2025.104405

[10] Yan, F., Chen, Y., Xia, Y., Wang, Z. and Xiao, R. (2023) An Explainable Brain Tumor Detection Framework for MRI Analysis. *Applied Sciences*, **13**, Article No. 3438. https://doi.org/10.3390/app13063438

[11] Gundogan, E. (2025) A Novel Hybrid Deep Learning Model Enhanced with Explainable AI for Brain Tumor Multi-Classification from MRI Images. *Applied Sciences*, **15**, Article No. 5412. https://doi.org/10.3390/app15105412

[12] Srinivas, V.R. and Parvathi, R. (2026) Explainable AI-Driven MRI-Based Brain Tumor Classification: A Novel Deep Learning Approach. *Frontiers in Artificial Intelligence*, **8**, Article ID: 1700214. https://doi.org/10.3389/frai.2025.1700214

[13] Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D. and Batra, D. (2017) Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. 2017 *IEEE International Conference on Computer Vision* (*ICCV*), Venice, 22-29 October 2017, 618-626. https://doi.org/10.1109/iccv.2017.74

[14] Chattopadhyay, A., Sarkar, A., Howlader, P. and Balasubramanian, V.N. (2017) Grad-CAM++: Improved Visual Explanations for Deep Convolutional Networks.

[15] Ribeiro, M.T., Singh, S. and Guestrin, C. (2016) "Why Should I Trust You?": Explaining the Predictions of Any Classifier. *Proceedings of the* 22*nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Francisco, 13-17 August 2016, 1135-1144. https://doi.org/10.1145/2939672.2939778

[16] Lundberg, S. and Lee, S.I. (2017) A Unified Approach to Interpreting Model Predictions.

[17] Zeiler, M.D. and Fergus, R. (2014) Visualizing and Understanding Convolutional Networks. In: Fleet, D., *et al*., Eds., *Computer Vision—ECCV* 2014, Springer International Publishing, 818-833. https://doi.org/10.1007/978-3-319-10590-1_53

[18] Iftikhar, S., Anjum, N., Siddiqui, A.B., Ur Rehman, M. and Ramzan, N. (2025) Explainable CNN for Brain Tumor Detection and Classification through XAI Based Key Features Identification. *Brain Informatics*, **12**, Article No. 10. https://doi.org/10.1186/s40708-025-00257-y

[19] Lecun, Y., Bottou, L., Bengio, Y. and Haffner, P. (1998) Gradient-Based Learning Applied to Document Recognition. *Proceedings of the IEEE*, **86**, 2278-2324. https://doi.org/10.1109/5.726791

[20] Ning, Y., Liu, J., Qin, L., *et al*. (2023) A Novel Approach for Auto-Formulation of Optimization Problems.

[21] Rastogi, D., Johri, P., Donelli, M., Agarwal, T., Tiwari, S. and Singh, P. (2025) XAI-BT-EdgeNet: Explainable Edge-Aware Deep Learning with Squeeze-and-Excitation for Brain Tumor Detection and Prediction. *Frontiers in Artificial Intelligence*, **8**, Article ID: 1676524. https://doi.org/10.3389/frai.2025.1676524

[22] Singh, R., Gupta, S., Ibrahim, A.O., Gabralla, L.A., Bharany, S., Rehman, A.U., *et al*. (2025) Advanced Dynamic Ensemble Framework with Explainability Driven Insights for Precision Brain Tumor Classification across Datasets. *Scientific Reports*, **15**, Article No. 29090. https://doi.org/10.1038/s41598-025-14917-w

[23] Hafeez, Y., Memon, K., AL-Quraishi, M.S., Yahya, N., Elferik, S. and Ali, S.S.A. (2025) Explainable AI in Diagnostic Radiology for Neurological Disorders: A Systematic Review, and What Doctors Think about It. *Diagnostics*, **15**, Article No. 168. https://doi.org/10.3390/diagnostics15020168

[24] Yoon, H.C. and Lin, L.P. (2025) Brain Tumor Classification in MRI: Insights from LIME and Grad-CAM Explainable AI Techniques. *IEEE Access*, **13**, 154172-154202. https://doi.org/10.1109/access.2025.3603272

[25] Gharaibeh, N. (2025) Enhancing Interpretability in Brain Tumor Detection: Leveraging Grad-CAM and SHAP for Explainable AI in MRI-Based Cancer Diagnosis. *Applied Computer Science*, **21**, 182-197. https://doi.org/10.35784/acs_7375

[26] Muhammad, D. and Bendechache, M. (2024) Unveiling the Black Box: A Systematic Review of Explainable Artificial Intelligence in Medical Image Analysis. *Computational and Structural Biotechnology Journal*, **24**, 542-560. https://doi.org/10.1016/j.csbj.2024.08.005

[27] Doshi-Velez, F. and Kim, B. (2017) Towards a Rigorous Science of Interpretable Machine Learning.

[28] Wang, D. and Tanaka, T. (2020) Robust Kernel Principal Component Analysis with $\ell_{2,1}$-Regularized Loss Minimization. *IEEE Access*, **8**, 81864-81875. https://doi.org/10.1109/access.2020.2990493

[29] McKinney, S.M., Sieniek, M., Godbole, V., Godwin, J., Antropova, N., Ashrafian, H., *et al*. (2020) International Evaluation of an AI System for Breast Cancer Screening. *Nature*, **577**, 89-94. https://doi.org/10.1038/s41586-019-1799-6

[30] Böhle, M., Eitel, F., Weygandt, M. and Ritter, K. (2019) Layer-Wise Relevance Propagation for Explaining Deep Neural Network Decisions in MRI-Based Alzheimer's Disease Classification. *Frontiers in Aging Neuroscience*, **11**, Article No. 194. https://doi.org/10.3389/fnagi.2019.00194