Article

Not peer-reviewed version

Bridging the Gap: Exploring Interpretability in Deep Learning Models for Brain Tumor Classification from MRI Images

Wandile Nhlapho , Marcellin Atemkeng * , Yusuf Brima , Jean-Claude Ndogmo *

Posted Date: 19 February 2024

doi: 10.20944/preprints202402.0960.v1

Keywords: transfer learning; deep learning; brain tumor classification; explanability and interpretability; Grad-Cam++; integrated gradients

Article

Bridging the Gap: Exploring Interpretability in Deep Learning Models for Brain Tumor Classification from MRI Images

Wandile Nhlapho [1], Marcellin Atemkeng [2,*], Yusuf Brima [3] and Jean-Claude Ndogmo [1,*]

[1] Department of Mathematics and Applied Mathematics University of Venda, Thohoyandou 0950, South Africa
[2] Department of Mathematics, Rhodes University, Grahamstown, 6139, Eastern Cape, South Africa.
[3] Computer Vision, Institute of Cognitive Science, Osnabrück University, Osnabrück, D-49090, Lower Saxony, Germany.
* Correspondence: m.atemkeng@gmail.com (M.A.); jean-claude.ndogmo@univen.ac.za (J.C.N.)

Abstract: The advent of deep learning (DL) has revolutionized medical imaging, offering unprecedented avenues for accurate disease classification and diagnosis. DL models have shown remarkable promise for classifying brain tumor from Magnetic Resonance Imaging (MRI) scans. However, despite their impressive performance, the opaque nature of DL models poses challenges in understanding their decision-making processes, particularly crucial in medical contexts where interpretability is essential. This paper explores the intersection of medical image analysis and DL interpretability, aiming to elucidate the decision-making rationale of DL models in brain tumor classification. Leveraging state-of-the-art DL frameworks with transfer learning, we conduct a comprehensive evaluation encompassing both classification accuracy and interpretability. Using state-of-the-art DL frameworks with transfer learning, we conduct a thorough evaluation covering both classification accuracy and interpretability. We employ adaptive path-based techniques to understand the underlying decision-making mechanisms of these models. Grad-CAM and Grad-CAM++ highlight critical image regions where the tumors are located.

Keywords: transfer learning; deep learning; brain tumor classification; explainability and interpretability; Grad-Cam++; integrated gradient

1. Introduction

The field of medical imaging has experienced a transformative paradigm shift with the emergence of deep learning (DL), unlocking unprecedented opportunities for accurate disease classification and diagnosis as indicated in [1–5]. In the area of brain tumor classification using Magnetic Resonance Imaging (MRI) scan images, these DL models have shown exceptional promise [1,2]. The significance of brain tumor diagnosis from MRI images lies in its pivotal role in modern healthcare, offering opportunities for early detection and timely treatment, as evident in [2,3,6]. DL models have shown remarkable potential in automating this process, yet their effectiveness and reliability in real-world medical applications, particularly in brain tumor classification, remain under scrutiny. The primary challenge stems from the need to achieve both *high accuracy* and *model interpretability*, as emphasized in [7].

Despite the impressive classification results produced by DL models, their "black-box" nature impedes a comprehensive understanding of the decision-making rationale. Also, their incorporation into clinical practice and decision-making needs adaptability, not just in terms of classification performance, but also in interpretability and explainability. This study investigates the junction of medical image analysis and DL interpretability. Brain tumors, characterized by their diversity and clinical significance as highlighted in [2,3], demand precise classification for tailored treatment plans and improved patient outcomes. While DL models have exhibited substantial prowess in this domain, their inherent opacity presents a challenge, as discussed in [3,6]. The central issue revolves around

 doi:10.20944/preprints202402.0960.v1

understanding how these models reach their conclusions, particularly crucial in the context of medical decision-making. The decision-making processes of DL models in medical imaging are currently under scrutiny through interpretability techniques, shedding light on how algorithms make predictions. The transparency and reliability of AI-driven medical diagnoses are augmented by these techniques, facilitating the identification of pertinent features and areas within images contributing to the model's choices.

To address this challenge, we conduct a broad performance evaluation of various state-of-the-art DL frameworks with transfer learning for brain tumor MRI classification. This evaluation encompasses not only the assessment of classification accuracy but also a focus on evaluating the interpretability of the models. Adaptive path-based techniques, including Gradient-weighted Class Activation Mapping (Grad-Cam), Grad-Cam++, Integrated Gradient (IG), and Saliency Mapping, are leveraged to unravel the underlying controlling processes of these models. The ultimate objective is to enhance the clinical utility of DL based on brain tumor identification by providing more transparent and interpretable results.

Techniques for interpretability, such as Grad-Cam++ [8], Grad-CAM [9], Integrated Gradients [10], and SHAP [11], have distinguished themselves by highlighting the image regions most important to the model's conclusions. For instance, Grad-CAM creates heatmaps that highlight significant image regions, aiding doctors in comprehending the model's focal point. Although DL models for diagnosing brain tumors have attained amazing levels of accuracy [12,13], their intricate internal mechanisms raise questions about their dependability and generalizability. To solve this, academics have started looking into how interpretable these models are. To improve a DL model's transparency, [14] created a brand-new technique for viewing tumor localization. It is difficult to gauge how well DL models can be interpreted. According to [15,16], several criteria have been put out to evaluate consistency, localization accuracy, and sensitivity to perturbations. These measures are intended to offer a consistent framework for evaluating how closely the model's explanations match clinical expectations and actual tumor sites. Medical practitioner's ability to make decisions may be improved by combining interpretable DL models into clinical practice. Recent research [17] has shown how radiologists and doctors can use interpretable model explanations to validate the model's predictions and produce more accurate diagnoses. Radiologists participated in the study in [18–20] to assess the interpretability of DL models in medical imaging. To get professional opinions on the value and comprehensibility of the model explanations, the study employed user studies. The research underlined how crucial interpretability is to fostering communication among medical specialists.

The exploration of multiple DL architectures holds considerable importance in evaluating their effectiveness within the specialized domain of brain tumor classification using MRI images. We conduct an ultra-careful analysis, each model, including AlexNet, DenseNet121, EfficientB0, GoogleNet, Inception V3, ResNet50, VGG16, VGG19, ViT Transformer, and Xception, is identified by a unique set of parameters and indicators, contributing to a degree of understanding of their merits. The evaluation criteria, spanning accuracy, precision, recall, and F1 score percentages, collectively offer a comprehensive and insightful depiction of the capabilities exhibited by these models. To further light up our findings, we implement interpretability techniques, such as Grad-CAM, Grad-CAM++, Integrated Gradient, and Saliency Mapping, unveiling the intricate decision-making processes and providing valuable insights into the regions and features significantly contributing to the models' classifications. This integrated approach not only gauges the models' performance but also enhances the transparency and interpretability of their decision-making procedure.

The rest of this paper is structured as follows. Section 2 discusses the literature review on explainability in medical imaging, while Section 3 covers the attribution methods used in this work and the transfer learning models. Section 4 presents the dataset and our proposed method. In Section 5, experimental results are presented, and conclusions are reported in Section 6.

 doi:10.20944/preprints202402.0960.v1

2. Literature on DL Models Explainability in Medical Imaging

The input image is passed through the algorithm via one forward as well as backward propagation. The resulting score is then computed forward, and the dependency gradient amongst the convolution layers is then determined to build an attribution map. This process is known as the Vanilla Gradient explainability method. It is an easy-to-understand attribution approach with little processing power requirements because of its simplicity. Vanilla Gradient and other attribution-based graphic aids for MRI imaging of brain tumors were evaluated using an attribution-based scheme named NeuroXAI [21]. Both feature segmentation and classification were visualized using these techniques. Vanilla Gradient produced noisier attribution maps than the other attribution techniques and had gradient saturation, which means that a change in a neuron has no effect on the network's output and so cannot be assessed. Similar results were observed utilizing Vanilla Gradient for feature visualization in [22], where the dissimilitude intensification juncture from computed tomography (CT) images is anticipated. Furthermore, Vanilla Gradient is unable to distinguish between classes (such as healthy and diseased) [23]. This demonstrates that Vanilla Gradient is unable to produce attribution maps that are distinct based on class. The deconvolution network (DeconvNet) is essentially comparable to Vanilla Gradient, with the primary distinction lying in the computation of gradients over a Rectified Linear Unit (ReLU) function [24].

The human brain's artery segmentation was studied using different attribution techniques for interlayer CNN visualization using TorchEsegeta, a structure for image-based DL algorithms that can be understood and explained [25]. Since other techniques also indicated non-vessel activation, their primary focus was on the vessels, Vanilla Gradient, and DeconvNET produced results that were more comprehensible to humans than other attribution techniques like Deep Learning Important Features (DeepLIFT) and Gradient-weighted Class Activation Mapping (GradCAM++).

Guided back propagation (GBP) integrates both the deconvNET [26] and the Vanilla Gradient. Comparing this approach to applying each technique separately yields less noisy attribution maps as there are fewer active voxels. As compared to Vanilla Gradient, GBP presented purpose-specific attribution maps in the NeuroXAI framework with a lot less noise [21]. An extra refining mechanism for the GBP was suggested in [26] to further reduce the quantity of noise and the influence of indiscriminate attributions on predicting brain disorders using MRI. GBP is likely to offer attribution maps with reduced noise, but it could furthermore produce attribution maps that are too sparse, which are unhelpful for comprehensive image characterization [27]. Although they are not class discriminative, all three of the gradient-based techniques are quite sensitive to the way the neural network layers collect information. ReLU and pooling layers may also cause local gradients to saturate. As a result, significant characteristics may disappear as the network's layers advance, which might lead to an inadequate model clarification or even a concentration on unimportant features. Layer-wise relevance propagation (LRP) is an Explainable Artificial Intelligence (XAI) technique that employs principles unique to LRP to propagate the class score backward through the neural layers to the input image [28]. The core idea of LRP is to preserve inter-neuron interdependence, ensuring that information acquired by one layer of neurons is equally transferred to the next lower layer. LRP addresses the challenges posed by the saturation problem since the decomposition relies on propagating relevance scores between neurons rather than gradients. In a study focused on the detection of abdominal aortic aneurysms using CT images, LRP demonstrated a distinct class difference based on activation differences in the aortic lumen [29].

Deep Learning Important Features (DeepLIFT) is an XAI technique that addresses the saturation problem by employing a neutral reference activation, such as the neuron activation in computed tomography (CT) scans without pathology or disease [30]. The difference between a new neuron's activation and the reference activation is described by this reference activation. An attribution map is generated by computing the contribution scores for each neuron based on these differences. To discern individuals with Multiple Sclerosis (MS) using MRI, DeepLIFT was compared to LRP and Vanilla Gradient [31]. Based on the quantitative evaluation, DeepLIFT extracts target-specific characteristics

 doi:10.20944/preprints202402.0960.v1

much better than Vanilla Gradient and marginally better than LRP. Gradient saturation is something that both LRP and DeepLIFT can handle, which might be why it outperforms Vanilla Gradient in this classification challenge. Among the most popular model-specific attribution techniques is the class activation map (CAM) [32,33]. Rather than using numerous dense layers, it employs a Global Average Pooling (GAP) layer, which adds linearity before the last dense layer and after the last convolution layer. Low-dimension attribution maps are produced by CAM since they only utilize information from the last convolution layer. As a result, the low-dimension CAM can show if a model can generally focus on particular targets, but because of its poor specificity, it is unable to discriminatively define characteristics depending on class [34,35]. Additionally, it was revealed through perturbation analysis of several attribution methodologies that gradient-based approaches have a better rigor than CAM [22]. However, CAM can be indicative when doing patch-based (more targeted) tumor analysis as opposed to whole image tumor analysis [36,37], or when the classes in a classification task exhibit obvious visual distinctions, such as between healthy and Alzheimer's brains [38].

The use of XAI techniques has increased dramatically as a result of COVID-19 detection [39]. Generally speaking, these techniques may be distinguished by either using the entire CT scan or only the lung segmentation for COVID-19 identification. In particular, there was a significant performance difference in attribution mapping for COVID-19 detection based on the entire picture. The most common attribution technique was Grad-CAM, an extension of CAM, which produced both very specific [40,41] and non-specific attributions [31,42,43], but generally was able to approximately pinpoint the possible COVID-19 lesions to produce reliable predictions. A priori segmentation of the lungs was proposed to eliminate the impact of non-target specific characteristics [43–49]. In this manner, only characteristics from the lungs may be extracted by both the DL algorithms and the XAI approaches. In this sense, the XAI techniques and the DL algorithms can only extract characteristics from the lungs. Compared to utilizing the whole CT image with Grad-CAM, this anatomically based XAI approach demonstrated greater specificity. This indicates the benefits of medical-based data reduction for DL and XAI algorithms, that is, lowering the number of trainable characteristics and/or getting rid of uninformative characteristics based on the input image. When the entire image was used, comparable non-target relevant attribution maps were observed as well (in the absence of data reduction) for cerebral hemorrhage detection [50], and automated grading of expanded perivascular spaces in acute stroke [51]. In a manner akin to the COVID-19 investigations, prior anatomical segmentation was employed to categorize and illustrate mortality risks based on cardiac PET [52], Alzheimer's disease [53] and schizophrenia based on MRI [54]. Grad-CAM's low-dimensional attribution maps, however, continue to cause poor specificity even while data handling reduces the prevalence of non-target specific characteristics [55,56]. The authors proposed that the active characteristics surrounding the tumor correlate to areas harboring occult microscopic illness based on the Grad-CAM attribution maps, in research for the categorization of lung cancer histology based on CT images [7]. It is more plausible, though, due to Grad-CAM's poor dimensionality, as CT lacks the spatial resolution necessary to identify these tiny illnesses.

In classification tasks when there is a discernible radiological difference between the classes, Grad-CAM, like CAM, can be class discriminative [10,57–60]. However, additional attribution methods like Vanilla Gradient and GBP should be utilized in cases of tasks with less clear radiological distinctions, such as predicting survival based on tumor feature, where Grad-CAM lacks fine-grained information [22,25]. In MRI imaging of brain tumors, research that paired GBP with Grad-CAM, a technique known as guided Grad-CAM (gGrad-CAM) showed improved localized attribution maps with greater resolution [21]. These supports integrate the benefits of attribution techniques for accurate and comprehensible model visualization. Numerous further enhanced versions of Grad-CAM, including Grad-CAM++, have been developed. To improve target-specific feature localization over Grad-CAM, Grad-CAM++ was introduced [6]. Grad-CAM may reduce the disparity in relevance between the various gradients since it averages the feature map gradients. Grad-CAM++ substitutes a weighted average, which quantifies each feature map unit's significance, in its stead. In terms of

 doi:10.20944/preprints202402.0960.v1

knee osteoarthritis prediction using MRI, it demonstrated better target-specific attribution maps than Grad-CAM [9].

3. Attribution Methods and Transfer Learning Models

We explore the complexities of transfer learning models and attribution methods. This investigation is essential because it offers a thorough grasp of the theoretical underpinnings of the models we trained and the attribution techniques used to ensure explainability. Through providing in-depth analyses of these techniques, we want to provide clarification on the fundamental ideas that underpin how they work. This theoretical foundation is essential to understanding the subtleties of the models and their interpretability, which is consistent with our goal of developing a deeper grasp of the theoretical foundations of these approaches in addition to putting them into practice.

3.1. Attribution Methods

We shall harness the power of Grad-Cam [9], Grad-Cam++ [8], Integrated Gradient [10], and Saliency Mapping [61], four ingenious path-oriented techniques, to compute the visual interpretability measure. These methods graciously provide a cartographic representation of the regions within the image that influenced the DL model's decision-making process.

3.1.1. Grad-Cam and Grad-Cam++

Grad-Cam and Grad-Cam++ focus on visualizing the areas in the image that contribute significantly to CNN's decision-making process, providing interpretability to the model's predictions. For Grad-Cam, let A be a feature map of class c, and Y^c the output that corresponds to class c. Its weights α_k^c at the ij position of the k-th feature map is expressed as:

$$\alpha_k^c = \frac{1}{Z} \sum_i \sum_j \frac{\partial Y^c}{\partial A_{ij}^k}, \tag{1}$$

where Z represents the feature map's size, and A_{ij}^k the activation of the unit in position ij of the k-th feature map. In Grad-Cam++, the gradient weight α_{ij}^{kc} and $ReLU$ function that correspond to class c are added and the weight w_k^c is expressed as:

$$w_k^c = \sum_i \sum_j \alpha_{ij}^{kc} \cdot ReLU\left(\frac{\partial Y^c}{\partial A_{ij}^k}\right). \tag{2}$$

Lastly, the localization heatmap L of class c in position ij is expressed as:

$$L_{ij}^c = \sum_k w_k^c \cdot A_{ij}^k. \tag{3}$$

$ReLU$ is applied to enhance the relevance of positive gradients. These equations (1),(2),(3) collectively represent the process of generating a heatmap that highlights the important regions in the input image for making predictions related to class c.

3.1.2. Integrated Gradient (IG)

Using this method, an attribution map is produced that shows the image parts that are important for the categorization choice. The output $h_c(y)$ represents the confidence score for predicting class c given a classifier h, input y, and class c. To compute Integrated Gradients (IG), we perform a line integral between a reference point y' and an image y in the vector field generated by the gradient of

 doi:10.20944/preprints202402.0960.v1

$h_c(y)$ with respect to the input space. This vector field helps IG determine the importance or attribution for each feature, such as a pixel in an image. Formally, IG is defined as follows for each feature i:

$$I_i^{IG}(y) = \int_0^1 \frac{\partial h_c(\gamma^{IG}(\alpha))\partial \gamma_i^{IG}(\alpha)}{\partial \gamma_i^{IG}(\alpha)\partial \alpha} d\alpha, \tag{4}$$

where $\gamma^{IG}(\alpha), \alpha \in [0,1]$ is the parametric function representing the path from y' to y, with $\gamma^{IG}(0) = y'$ and $\gamma^{IG}(1) = y$. Specifically, γ^{IG} is a straight line connecting y' and y.

3.1.3. Saliency Mapping

A method called saliency mapping uses an analysis of each pixel's impact on the classification score to determine how salient it is in an input image. Saliency maps show the visual regions that influence the classification choice with a linear scoring model [61]:

$$S_c(I) = w_c^T I + b_c, \tag{5}$$

where b_c is the bias for class c, w_c is the vector of weights, and I is a single-dimensional vectorized description of the image's pixels. It is easy to see that in such a situation, the related pixels of I are important based on the importance of the components of w_c. We cannot simply use this insight since $S_c(I)$ is a non-linear function of the image in deep convolutional networks. However, by evaluating the first-order Taylor expansion, we can roughly estimate the class score function with a linear function in the vicinity of a given image I_0:

$$S_c(I) \approx w^T I + b, \tag{6}$$

where w is the derivative of S_c with respect to the image I at the point (image) I_0:

$$w = \left.\frac{\partial S_c}{\partial I}\right|_{I_0}. \tag{7}$$

Equation (7) can also be used to calculate the image-specific class saliency. In this case, the derivative's magnitude shows which pixels require modification.

3.2. Transfer Learning Models

Transfer Learning (TL) is conceptualized as providing a head start to a model. Consider a pre-trained model that excels in a specific task. When faced with having the model excel in a related yet slightly different challenge, TL becomes pertinent. Rather than training the model from the initial stage, the approach involves leveraging the acquired knowledge, specifically the parameters, from the initial task and applying it to the new, related task. This process facilitates a more efficient adaptation of the model to the novel task by capitalizing on the previously acquired expertise. The TL models under consideration include AlexNet, ResNet50, DenseNet121, GoogleNet, Xception, Inception V3, VGG16, VGG19, Vision Transformer (ViT), and EfficientNetB0. These models are applied to the classification task we aim to examine. Each architecture has unique characteristics, and we will explore the learning capabilities inherent in each of them, assessing their performance and suitability for the specific classification task at hand.

3.2.1. AlexNet

Deep CNN architecture AlexNet made significant contributions to the development of DL and computer vision. By winning the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) in 2012, it was created by [62] and considerably advanced the state-of-the-art in picture/image classification problems. The success of AlexNet paved the path for more intricate convolutional neural network architectures and catalyzed numerous other developments in DL for computer vision. Modern neural

 doi:10.20944/preprints202402.0960.v1

network models are still influenced by their design tenets, such as the usage of convolutional layers, ReLU activations, and dropout regularization. It contains 3 fully connected layers and 5 convolutional layers for a total of 8 layers. The convolutional layers are separated by max-pooling layers, which are used to down-sample and capture important information.

3.2.2. ResNet50

The residual network ResNet50 is 50 layers deep [63]. The ResNet50 design integrates a combination of convolution filters of various sizes to address the degeneration of CNN models and shorten training times. A max pool layer, an average pool layer, and a total of 48 convolutional layers make up this architecture.

3.2.3. DenseNet121

The DenseNet model was proposed in [65]. Its primary components are DenseBlock (DB), transition layer, and growth rate. DenseNet121 has the advantage of requiring fewer parameters, allowing for the training of deeper models during computation. Additionally, the fully connected layer of the model also has a regularization effect, which can help prevent overfitting on smaller datasets.

3.2.4. GoogleNet

The best-performing model, GoogLeNet, was presented by Google at the 2014 ImageNet Large-Scale Visual Recognition Challenge (ILSVRC14) [66]. The inner layers of the neural network were expanded to output diverse correlation distributions, based on the theory that achieving different probability distributions highly correlated with the input data would optimize the efficiency of each layer's neural network output.

3.2.5. Xception

Inception V3 was updated by Google to create Xception [67], which split regular convolution into spatial convolution and point-by-point convolution. Depthwise Separable Convolution (DSC) was used in place of the original Inception module. While point-wise convolution utilizes a 1-by-1 kernel to convolve point by point, reducing the number of parameters and computations, spatial convolution is conducted on each input channel.

3.2.6. Inception V3

Inception V1 was the name given to Google's initial proposal of GoogLeNets, which was followed by Inception V2 and Inception V3 [68] the following year. To reduce the size of feature maps, Inception V3 employs convolutional layers with a stride of 2 in combination with pooling layers. The first Inception module of Inception V3 replaces the 7 by 7 layer convolutional layer with 3 by 3 layer convolutional layers, which is a modification from Inception V2's first Inception module. The network's width and depth are increased in Inception V3 with the aforementioned upgrades to enhance performance.

3.2.7. VGG16

The contemporary transfer learning model VGG16, boasting sixteen weighted layers, stands as a state-of-the-art solution. Demonstrating its power on the ImageNet dataset, the model achieved an accuracy rate of 92.7% for the top five test results. The VGG16 got the top spot in the Large-Scale Visual Recognition Challenge (ILSVRC) organized by the Oxford Visual Geometry Group (VGG) [69]. The increased depth of the VGG model enables it to aid the kernel in capturing more intricate features.

 doi:10.20944/preprints202402.0960.v1

3.2.8. VGG19

Within the VGG19 model, an extension of the VGG16 architecture with 19 weighted layers, three additional fully connected (FC) layers contribute to a total of 4096, 4096, and 1000 neurons respectively, as reported in [69]. This model encompasses a Softmax classification layer alongside five Max pool layers. The convolutional layers within the architecture use the ReLU activation function.

3.2.9. EfficientNetB0

The EfficientNetB0 CNN architecture was proposed in [70]. Enhancing accuracy and efficiency through balanced scaling of the model's depth, breadth, and resolution is the aim of EfficientNet. The design presents a fixed ratio compound scaling technique that scales all three dimensions of depth, breadth, and resolution consistently.

3.2.10. Vision Transformer ViT

The architecture adopts the transformer's encoder component, revolutionizing image processing by segmenting the image into patches of a specified size like 16x16 or 32x32 dimensions [71,72]. This patch-based method enhances training with smaller patches. After flattening, patches are fed into the network. Unlike traditional neural networks, the model lacks positional information about the sequence of samples. To address this, the encoder incorporates trainable positional fixed vectors, eliminating the need for hard-coded positions.

3.3. Performance Evaluation

The accuracy and efficiency of deep learning models are evaluated in terms of how well they perform various tasks. Some of the most important performance measures frequently used are accuracy, precision, recall, and the so-called F_1 score. To describe these measures let us denote by TP, TN, FP, and FN the number of true positives, true negatives, false positives, and false negatives. Here, TP are instances where the model accurately predicted the presence of the positive class. For example, if the examination correctly identified 95 out of 100 patients with an ailment as afflicted with the ailment, those 95 are genuine positives; TN are instances where the model accurately predicted the absence of the negative class. If the examination correctly identified 80 out of 100 healthy individuals as not having the ailment, those 80 are genuine negatives; FP are instances where the model predicted a positive outcome when it ought to have been negative. If the examination mistakenly categorized 10 healthy individuals as having the ailment, those 10 are false positives and FN are instances where the model predicted a negative outcome when it ought to have been positive. If the examination overlooked 5 individuals with the ailment and labeled them as healthy, those 5 are false negatives. The performance measures are defined as:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{8}$$

$$Precision = \frac{TP}{TP + FP} \tag{9}$$

$$Recall = \frac{TP}{TP + FN} \tag{10}$$

$$F_1 = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall}. \tag{11}$$

 doi:10.20944/preprints202402.0960.v1

4. Data and Proposed Method

4.1. Dataset

The dataset used in this study was curated by Navoneel Chakrabarty and Swati Kanchan [73]. For easy access and reference, the dataset is currently available on Kaggle[1].

This brain tumor dataset comprises 3264 2D slices of T1-weighted contrast-enhanced images, encompassing three distinct types of brain tumors—glioma, meningioma, and pituitary tumors—along with images of a healthy brain. The dataset has been partitioned into 2937 images for training and 327 for testing purposes. Figure 1 visually presents samples from each of the four classes within the dataset and Figure 2 illustrates the proportion of each of the four classes in the dataset

Gliomas are tumors derived from glial cells and can manifest as either benign or malignant. Among them, glioblastoma multiforme stands out as a particularly aggressive variant, posing significant challenges in terms of therapeutic intervention [3]. Pituitary Tumors which arise in the pituitary gland, these tumors can disrupt hormonal balance. They may present as growths that secrete hormones or as non-functioning growths. Common sub-types include prolactinomas and growth hormone-secreting tumors, each with its distinct clinical implications. Meningiomas are generally benign, slow-growing tumors originating from the meninges. The symptoms associated with meningiomas vary based on the size and location of the tumor, making their clinical presentation diverse and often dependent on individual cases [3].

Understanding the unique features and characteristics of these brain tumors is essential for accurate diagnosis and effective treatment strategies. The use of this curated dataset allows for in-depth exploration and analysis of these distinct tumor types, contributing valuable insights to the field of medical imaging and diagnostics with deep learning and interpretability.

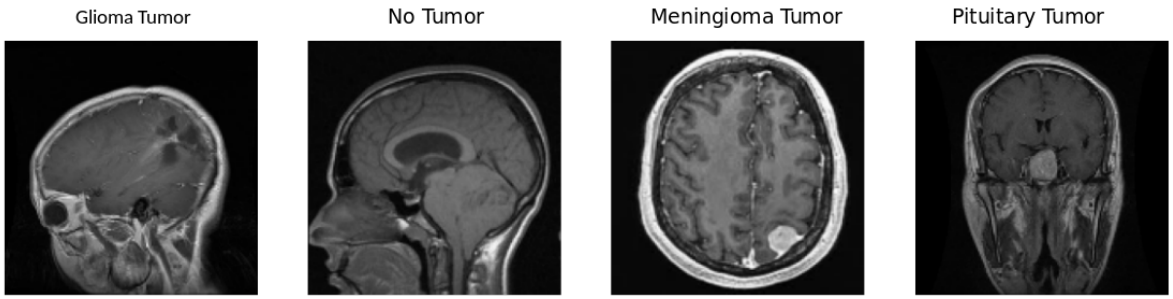

Figure 1. Sample of each image in the dataset. A Glioma Tumor image typically exhibits abnormal growth in the brain, indicating potential malignancy. No Tumor images represent a healthy state without any abnormal growth or lesions. Meningioma Tumor images showcase tumors arising from the meninges, the protective layers around the brain, and the spinal cord. Pituitary Tumor images depict tumors in the pituitary gland, influencing hormone regulation and potentially affecting various bodily functions.

[1] https://www.kaggle.com/datasets/sartajbhuvaji/brain-tumor-classification-mri/data

 doi:10.20944/preprints202402.0960.v1

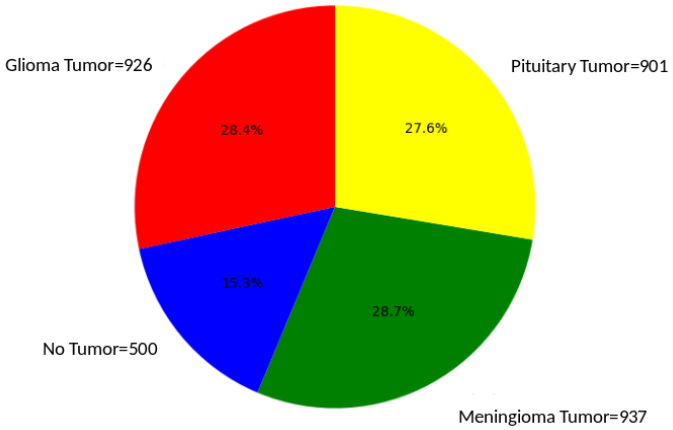

Figure 2. Brain tumor distribution. This pie chart effectively visualizes the relative proportions of different brain tumor types, offering a clear and concise representation of the distribution within the studied sample. The dataset comprises 926 MRI images of Glioma Tumors, 500 images with No Tumors, 901 images featuring Pituitary Tumors, and 937 images showing Meningioma Tumors.

4.2. Proposed Method

This paper proposes a comprehensive framework for brain tumor classification while integrating the model's explainability. This approach comprises two phases: phase (a) for classification and phase (b) for explainability, as illustrated in Figure 3.

- Phase (a): This phase integrates a CNN model with TL for the classification of brain tumor types based on MRI data. In this approach, the acquired features of a pre-trained CNN model serve as an initial foundation, proving particularly advantageous in scenarios involving sparsely labeled data. The data is fed into a CNN, which subsequently processes the data through a convolutional layer to capture intricate patterns and spatial hierarchies within the MRI images. Following this, the pooling layer is employed to down-sample and reduce the feature space, optimizing computational efficiency. Progressing along the activation path, the dense layer plays a pivotal role in transforming high-level features for effective classification. Finally, the model makes decisions about tumor types based on the combination of these learned features. When the decision is made, a medical expert becomes curious and seeks to understand how the model makes decisions based on their expertise.
- Phase (b): This phase uses explainability techniques, including Grad-Cam, Grad-Cam++, Integrated Gradient, and Saliency Mapping. The explanation aims to shed light on how the CNN model arrives at its classifications, providing valuable insights to the medical expert. Grad-Cam and Grad-Cam++ offer visualization of crucial regions in the MRI images that contribute to the model's decisions. Integrated Gradient provides a comprehensive understanding of feature importance by perturbing input features, while Saliency Mapping highlights salient features influencing the classification. Together, these explainability techniques bridge the gap between the model's predictions and human interpretability, fostering trust and comprehension in the application of DL models to medical imaging. After the model is explained, the focus shifts smoothly to the part where medical experts take over and make sense of it. The explained visualizations and insights provided by techniques like Grad-Cam, Grad-Cam++, Integrated Gradient, and Saliency Mapping serve as a bridge between the complex world of DL classifications and the expertise of medical experts. These outputs are presented to medical experts, allowing them to interpret and comprehend the rationale behind the model's decisions.

 doi:10.20944/preprints202402.0960.v1

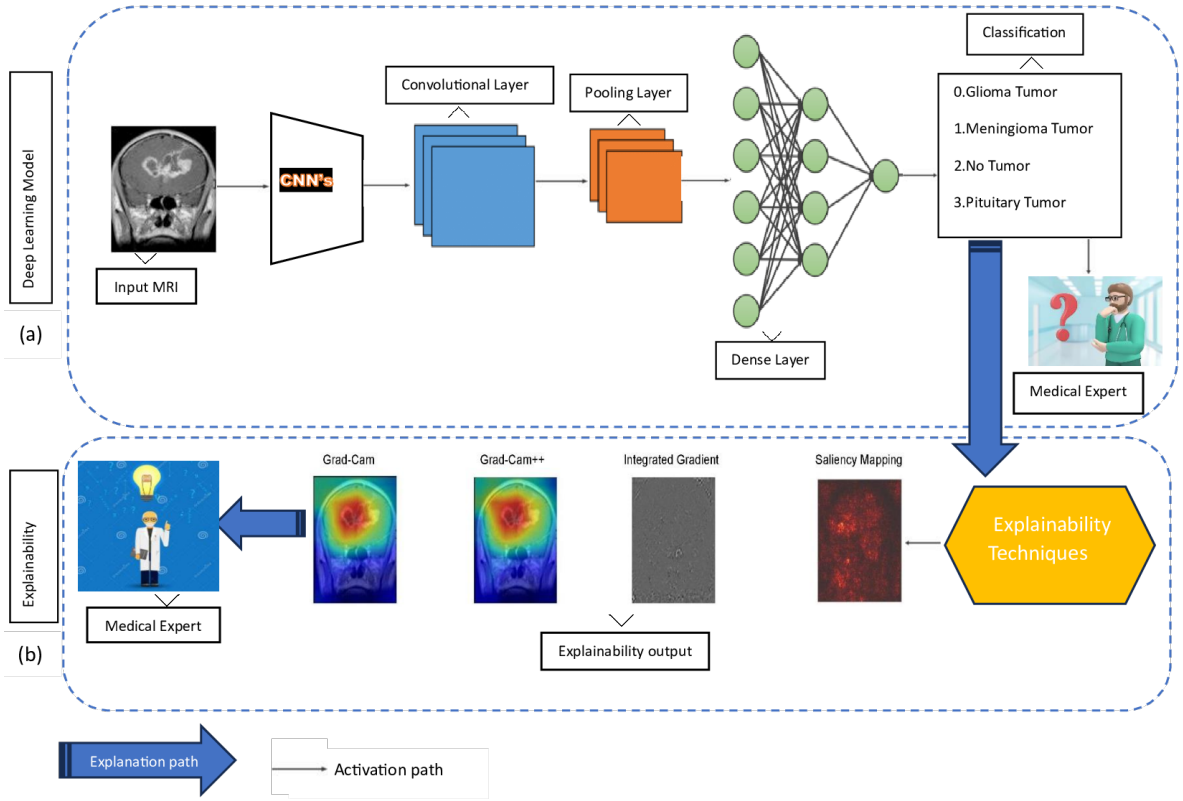

Figure 3. Comprehensive Framework for Brain Tumor Classification: Integrating Classification Accuracy and Model Explainability. (a) In the initial phase of our study, we focused on classifying different types of tumors. However, despite achieving classification results, we found ourselves in the dark about why the models made specific decisions. (b) To address this gap, we transitioned to the next stage of our investigation, delving into explainability methods. In this phase (b), we aim to shed light on the decision-making processes of our models. By employing explainability tools, we seek to unravel the factors and features that influenced the model's classifications. This shift allows us to not only categorize tumors accurately but also understand the underlying rationale behind the model's predictions, contributing to a more informed and interpretable outcome.

4.3. Training, Regularization, and Testing

This section elaborates on the model training process, hyperparameter tuning, and model testing, with a specific emphasis on the computational infrastructure used.

The models were trained using a DL approach, leveraging a CNN architecture with TL. Dataset was subjected to shuffling, ensuring a random order. Afterward, the dataset was divided, allocating 90 % for training purposes and reserving 10 % for validation. This approach aims to minimize bias and foster robustness in model training and evaluation. The dataset was split into training and validation sets, with the training set containing 2937 images and validation 327 images. The CNNs were implemented using Tensorflow version 2.12.0. The training process involved feeding the network with batches of images, where the model learned to identify patterns and features associated with each class of brain tumor. The choice of the loss function, optimizer (e.g., Adam, SGD), and learning rate played crucial roles in guiding the training process. These choices were fine-tuned based on the performance of the model on the validation set.

Hyperparameters, such as the number of layers, filter sizes, and dropout rates, were optimized to enhance the model's predictive capabilities. This involved a systematic exploration of different hyperparameter combinations using techniques like random search. The goal was to find the configuration that resulted in the best model performance as shown in Table 1. Additionally, TL

 doi:10.20944/preprints202402.0960.v1

was employed, using pre-trained models such as ResNet or VGG and fine-tuning them for the specific task of brain tumor classification.

The testing phase aimed to evaluate the trained model's generalization to new, unseen data. A separate test set, consisting of 327 images, was used for this purpose. Metrics such as accuracy, precision, recall, and the so-called F1 score were computed to assess the model's performance on each brain tumor class. The models were rigorously evaluated to ensure robustness and reliability in real-world scenarios. Careful attention was paid to handling imbalances in class distribution to avoid biased performance metrics.

The model training and hyperparameter tuning of our pre-trained models demanded substantial computational resources, and the efficiency of the process hinged on the specifications of our Dell computer, featuring an Intel Core i7 processor. The use of a powerful GPU, specifically the P100 model, played a pivotal role in accelerating the training speed. With 32 GB of system RAM, our system efficiently handled the computational load, facilitated by the Kaggle kernel environment. Implementation was carried out in Python, using Keras and TensorFlow version 2.12.0.

Table 1. Training Hyperparameters

Hyperparameter	Setting
Batch size	32
Learning rate	0.001
Epochs	10
Training split	90%
Test split	10%
Optimizer	Adam
Input size	150×150 pixels
Loss function	Categorical cross-entropy

5. Results

In this section, we present the outcomes of our study, offering a comprehensive overview of the key findings and their interpretation. Through this detailed analysis, we aim to provide a thorough understanding of the performance, interpretability, and decision-making processes inherent in the deep learning models employed in this work. Our contributions extend to the broader field of model interpretability, enhancing the collective knowledge within the research community.

5.1. Classification Results

We begin by discussing the training results, and offering insights into the model training process. The detailed training outcomes are presented in Table 2 and visually represented in Figure 4. Moving on to the test results, a summary is provided in Table 3, and Figure 5 visually highlights the superior performance of the best model.

This section also undertakes a thorough analysis and interpretation of the confusion matrix derived from the classification. A detailed examination follows, shedding light on the interpretability results. Specifically, results for adaptive path-based techniques, such as Grad-Cam, Grad-Cam++, Integrated Gradient, and Saliency mapping, are discussed in-depth. These techniques have enhanced our understanding of how deep learning models make classification decisions.

The training outcomes, as detailed in Table 2, offer insights into the model's ability to learn from the training data. High training accuracy and low training loss often signify successful training, yet these metrics may not necessarily ensure performance on the test data. Among the models investigated in this study, namely DenseNet121, EfficientB0, GoogleNet, Inception V3, ResNet50, and Xception, stand out with an acceptable training accuracy, ranging from 99.86% to 100%. These high accuracy scores indicate that these models have effectively learned the statistical regularities present in the training data. Furthermore, the associated training loss values are remarkably low, showcasing

 doi:10.20944/preprints202402.0960.v1

the models' efficiency in minimizing errors during the training phase. It is imperative to note that achieving high training accuracy does not necessarily guarantee superior performance on unseen datasets, emphasizing the importance of comprehensive evaluation of the test data for a more robust assessment of model generalization.

Table 2. Training Results

Model Name	Training Accuracy	Loss
AlexNet	0.8763	0.3233
DenseNet121	0.9986	0.0057
EfficientB0	0.9991	0.0042
GoogleNet	0.9997	0.0.0027
Inception V3	0.9989	0.0084
ResNet50	0.9991	0.0044
VGG16	0.8698	0.4011
VGG19	0.8570	0.3953
Vision Transformer	0.7484	0.5115
Xception	1.0000	0.0021

In contrast, models with notably lower training accuracy, spanning from 74.84% to 87.63%, such as AlexNet, VGG16, VGG19, and Vision Transformer, while still demonstrating commendable performance on the training data, exhibit comparatively higher training loss values. This suggests a slightly higher level of modeling error during the learning process for these models. The inferior performance of VGG16, VGG19, ViT Transformer, and AlexNet may be due to a combination of huge parameter counts and excessive model complexity, which may not be adequately aligned with the task's features. To increase the generalization capacity of these models, regularization strategies like dropout or batch normalization may need to be further refined or optimized.

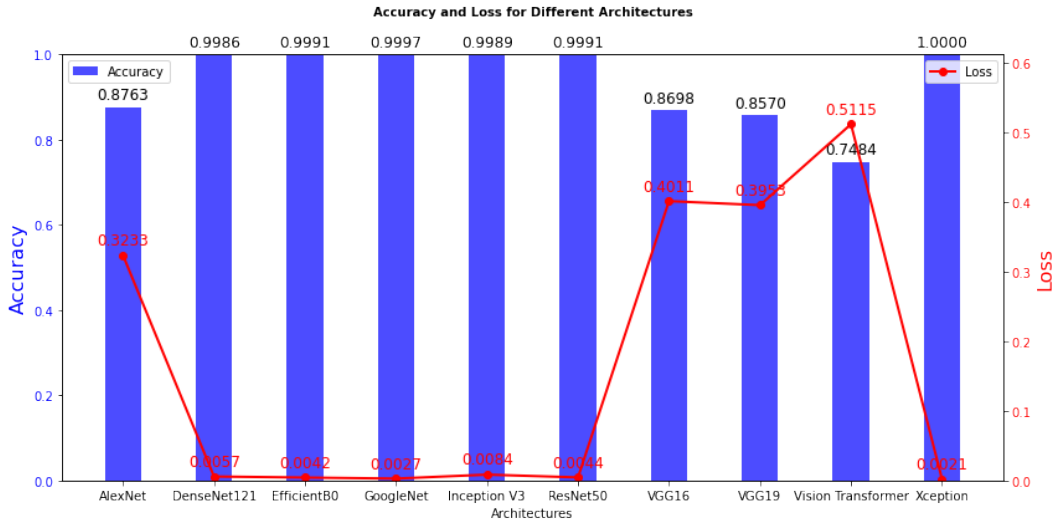

Figure 4. Training Accuracy and loss are the two main keys that are shown in this figure, which shows the training outcomes of several models. A distinct DL model is represented by each bar in the plot, and the height of the bar indicates the model's training accuracy. Additionally, the loss values for each model are displayed as a line plot superimposed on the same graph.

 doi:10.20944/preprints202402.0960.v1

Table 3. Test dataset classification results where various metrics are considered for each model.

Model Name	Parameters	Accuracy %	Precision %	Recall %	F1 Score %
AlexNet	61.9M	78	80	77	77
DenseNet121	8.1M	97	97	97	97
EfficientB0	5.3M	98	98	98	98
GoogleNet	11.2M	91	93	92	92
Inception V3	23.9M	96	97	96	96
ResNet50	25.6M	95	96	96	96
VGG16	138.4M	85	85	86	85
VGG19	143.7M	85	85	85	85
ViT Transformer	86M	70	72	72	70
Xception	22.9M	96	97	96	96

The classification results, as detailed in Table 3, provide valuable insights into the model's capability to classify the test data, it becomes evident that EfficientB0 emerges as the superior model among all those considered in this study. Upon closer examination of the results, it is apparent that some of the models may not be suited for this specific task, with some requiring more computational resources due to higher parameter counts, as observed in VGG16 and VGG19.

Figure 5 compares the 10 deep learning models used in this study, highlighting each model's performance across key criteria such as accuracy, precision, recall, and F1 score percentages. The models include AlexNet, DenseNet121, EfficientB0, GoogleNet, Inception V3, ResNet50, VGG16, VGG19, ViT Transformer, and Xception. This graphical representation offers a concise summary of these models' effectiveness in classifying brain tumors from MRI images, enabling a quick and informative comparison.

EfficientB0, exhibiting the best performance, shows great promise and warrants further consideration, particularly in medical imaging applications. The insights gained from this comprehensive evaluation contribute to the selection of an optimal deep learning model for the task of classifying brain tumors, emphasizing the potential impact of EfficientB0 in medical imaging.

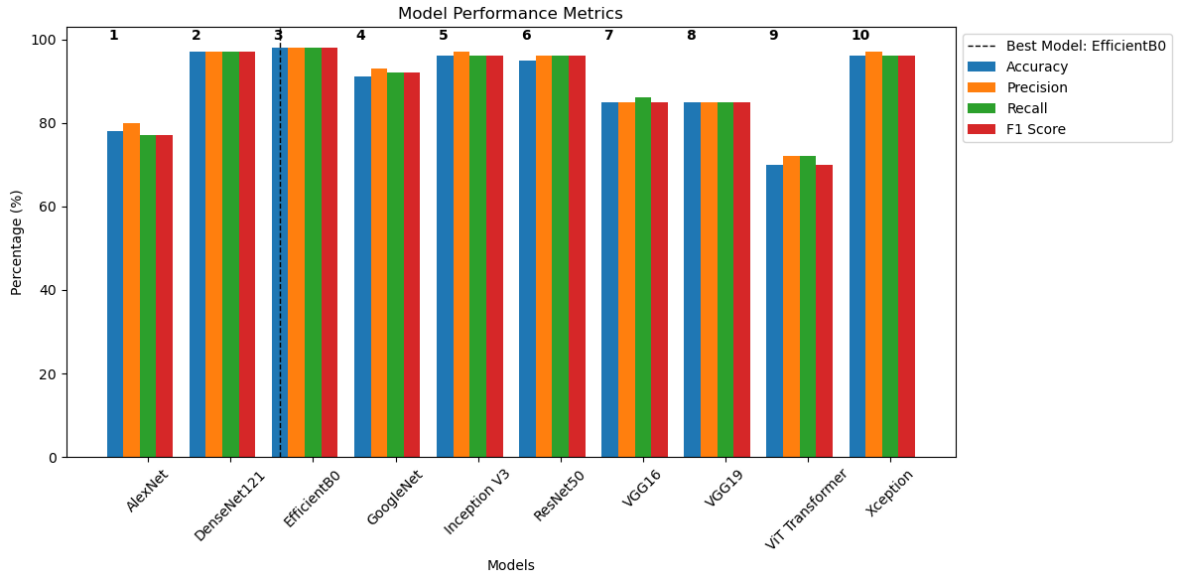

Figure 5. Visualizing the comparison of different DL models to provide a clear overview of their performance. The main performance indicators used to assess the effectiveness of various models are highlighted in the legend. The top model is named as EfficientNetB0 due to its excellent results in accuracy, precision, recall, and F1 score, which demonstrate its ability to provide well-balanced and accurate predictions on the provided test dataset.

 doi:10.20944/preprints202402.0960.v1

Figure 6 displays the confusion matrix for the EfficientNetB0 model. The diagonal elements represent samples that were accurately predicted. Out of the total 327 samples, 321 were predicted correctly, resulting in an overall accuracy of 98%. The element at row 1 and column 2, a value of 2, indicates that EfficientNetB0 incorrectly learned the classification boundary between classes 1 and 2. This implies that the model confused data initially belonging to class 2 with class 1. Conversely, the element at row 4 and column 1, a value of 0, signifies that the classification boundary between classes 1 and 4 was correctly learned by EfficientNetB0, and the model did not confuse data initially belonging to class 4 with class 1.

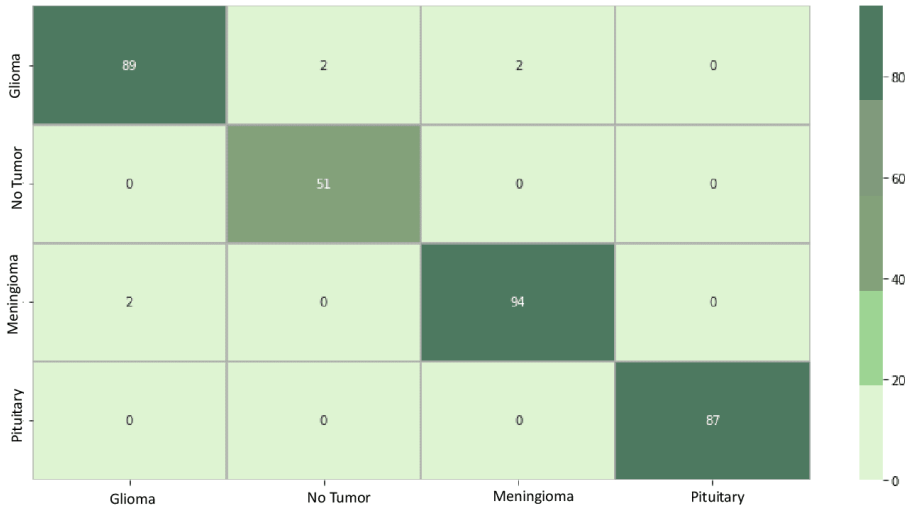

Figure 6. EfficientNetB0 confusion matrix. The matrix systematically breaks down the model's predictions, highlighting instances of true positives (correctly identified cases), true negatives (correctly rejected cases), false positives (incorrectly identified cases), and false negatives (missed cases) for each tumor class.

For model DenseNet121 in Figure 7, the confusion matrix shows that out of the total 327 samples, 320 were accurately predicted, resulting in an overall accuracy of 97%. The value 3 in row 1 and column 2 indicates that the classification boundary between classes 1 and 2 was incorrectly learned by DenseNet121, suggesting that the model confused data initially belonging to class 2 with class 1. Similarly, the presence of 0 in row 2 and column 1 implies that the classification boundary between classes 2 and 1 was correctly learned by DenseNet121, and the model did not confuse data initially belonging to class 1 with class 2.

 doi:10.20944/preprints202402.0960.v1

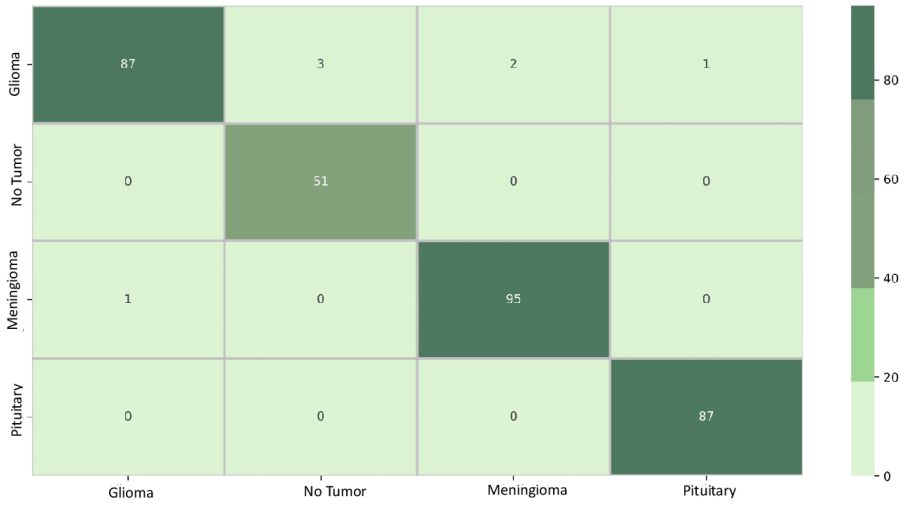

Figure 7. DenseNet121 confusion matrix.

Figure 8 depicts the confusion matrix for model Xception. Out of the total of 327 samples, 315 were accurately predicted, resulting in an overall accuracy of 96%. The value 1 in row 1 and column 2 indicates that the classification boundary between classes 1 and 2 was incorrectly learned by Xception, suggesting that the model confused data initially belonging to class 2 with class 1. Additionally, the presence of 0 in row 1 and column 4 signifies that the classification boundary between classes 1 and 4 was correctly learned by Xception, and the model did not confuse data initially belonging to class 4 with class 1.

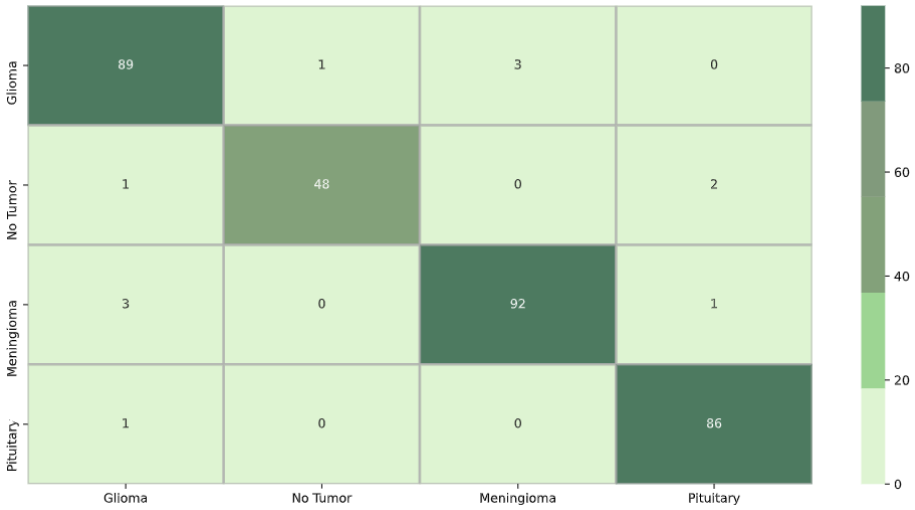

Figure 8. Xception confusion matrix.

5.2. Interpretability Results

In Figures 9, 10, and 11, we present a detailed look into the explainability results of our top-performing models: EfficientNetB0, DenseNet121, and Xception. Figures 9 provide insights into the interpretability of EfficientNetB0, highlighting crucial image regions using techniques such as GradCAM, GradCAM++, Integrated Gradients, and Saliency Mapping. Moving to Figures 10, we explore the explainability of DenseNet121, our 2nd best model, uncovering the significant features influencing its predictions. Figures 11 reveals the interpretability degrees of Xception, our 3rd best model, showcasing the impact of various image regions on classification decisions. These visualizations

 doi:10.20944/preprints202402.0960.v1

offer a transparent view into the decision-making processes of our models, facilitating understanding and trust.

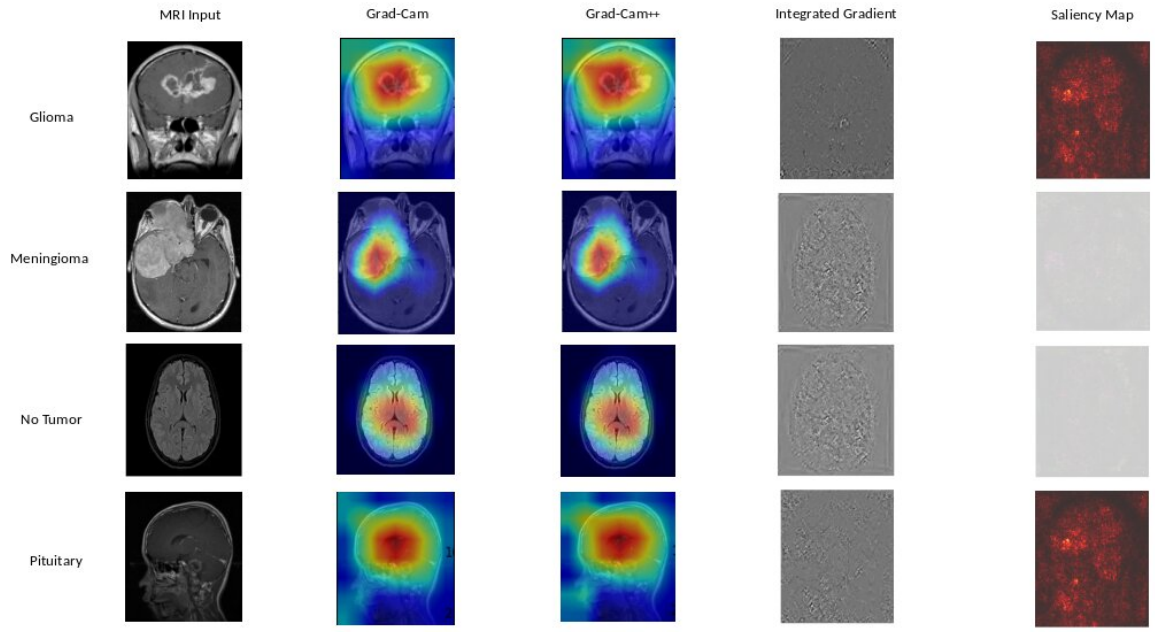

Figure 9. EfficientNetB0 explainability: We used a variety of explainability approaches, such as Grad-CAM, Grad-CAM++, Integrated Gradient, and Saliency Mapping, in our evaluation of EfficientNetB0 for brain tumor classification. These techniques played an important role in assisting in identifying the specific regions in MRI scan images that corresponded to the tumor types such as glioma, meningioma, no tumor, and pituitary. By using these techniques, we were able to determine the critical locations for the categorization of each tumor type and obtain important insights into EfficientNetB0's decision-making process.

 doi:10.20944/preprints202402.0960.v1

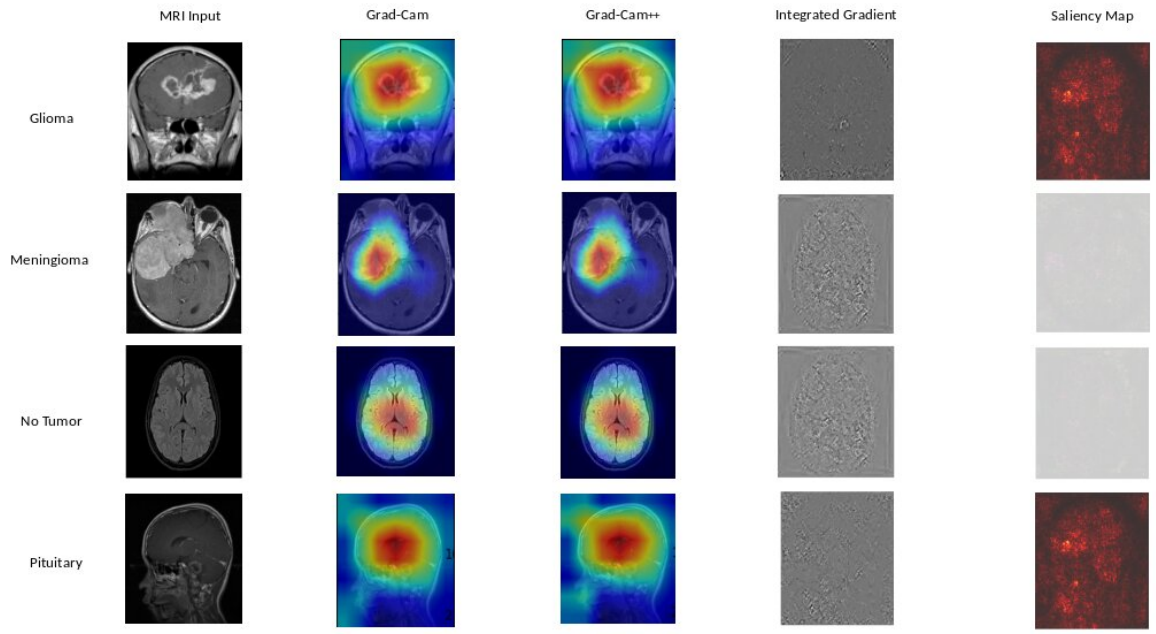

Figure 10. DenseNet121 explainability: We used a variety of explainability approaches, such as Grad-CAM, Grad-CAM++, Integrated Gradient, and Saliency Mapping, in our evaluation of DenseNet121 for brain tumor classification. These techniques played an important role in assisting in identifying the specific regions in MRI scan images that corresponded to the tumor types such as glioma, meningioma, no tumor, and pituitary. By using these techniques, we were able to determine the critical locations for the categorization of each tumor type and obtain important insights into DenseNet121's decision-making process.

 doi:10.20944/preprints202402.0960.v1

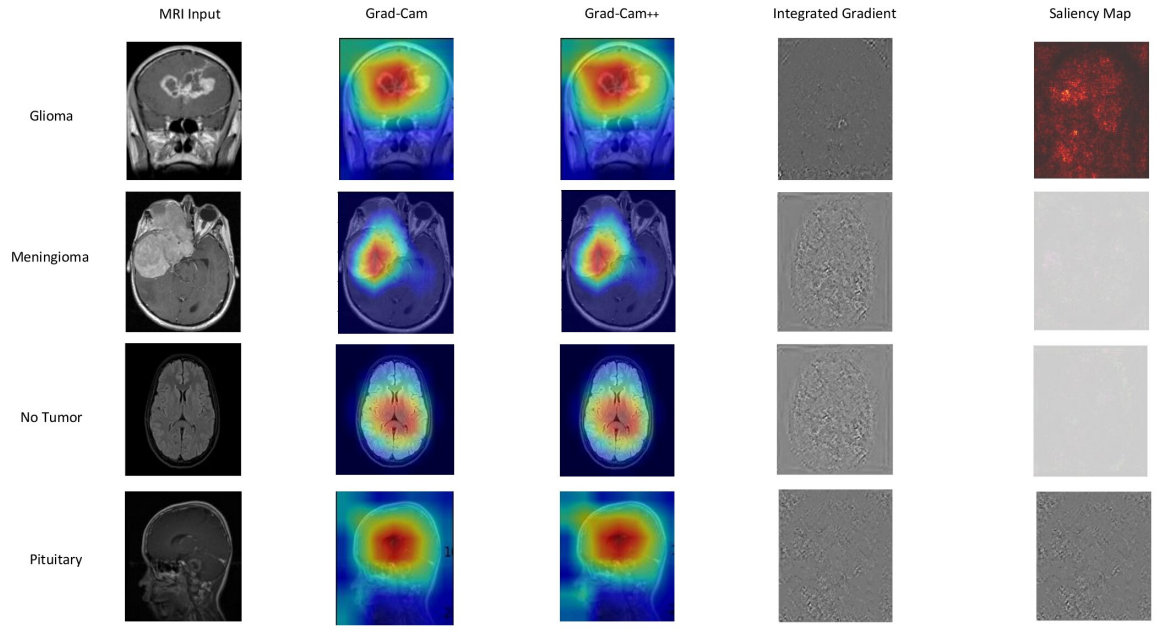

Figure 11. Xception explainability: We used a variety of explainability approaches, such as Grad-CAM, Grad-CAM++, Integrated Gradient, and Saliency Mapping, in our evaluation of Xception for brain tumor classification. These techniques played an important role in assisting in identifying the specific regions in MRI scan images that corresponded to the tumor types such as glioma, meningioma, no tumor, and pituitary. By using these techniques, we were able to determine the critical locations for the categorization of each tumor type and obtain important insights into Xception's decision-making process.

Based on our obtained results, both the Grad-Cam and Grad-Cam++ methods demonstrate similar outcomes, offering visual explanations for the decision-making processes of EfficientNetB0, DenseNet121, and Xception in predicting brain tumors. These methods accurately pinpoint the exact location of the tumor, and the visualizations generated by both approaches closely align, suggesting a consistent portrayal of the crucial regions or features influencing the model's predictions. For a more in-depth examination, refer to Figure 9 for insights into EfficientNetB0, Figure 10 for additional perspectives on DenseNet121, and for Xception, consult Figure 11. Our evaluation of visual explanations centers on their capacity to accurately describe the tumor's location within the brain, to foster confidence in human interpreters.

However, it is essential to note that Integrated Gradient faced challenges in precisely pinpointing the tumor's location, and Saliency Mapping exhibited some noise. Despite these challenges, both Grad-Cam and Grad-Cam++ consistently provided excellent visual explanations for the decision processes of EfficientNetB0, DenseNet121, and Xception, significantly enhancing our understanding of their predictive mechanisms.

5.3. Discussion

The findings of evaluating the performance of different DL models provide intriguing new information on the relationship between total accuracy and model complexity, which is reflected by the number of parameters shown in Table 3. DenseNet121 and EfficientB0 show impressive accuracy values of 97-98% with a lot fewer parameters, indicating how well they use information for the tasks at hand. On the other hand, accuracy is about 85% for VGG16 and VGG19, which have importantly greater parameter counts of more than 138 million. There is a clear exchange between model accuracy and complexity, highlighting the necessity of striking a balance. Interestingly, models with a reasonable amount of parameters, such as AlexNet, ResNet50, GoogleNet, and Xception, achieve competitive

 doi:10.20944/preprints202402.0960.v1

accuracy in the 80-96% range, striking a medium ground. Even with a low number of parameters, the ViT Transformer shows greater accuracy variability, highlighting the impact of architectural design on model performance. This study emphasizes how crucial it is to carefully weigh model complexity about the particular task at hand. Overly complicated models may not necessarily result in appreciably higher accuracy and may potentially be more prone to overfitting.

Grad-CAM, Grad-CAM++, Integrated Gradient, and Saliency Mapping show important areas that match with glioma, meningioma, and pituitary classifications. What stands out is the remarkable similarity in outcomes between Grad-CAM and Grad-CAM++ methods, precisely pinpointing tumor locations. This consistency paints a clear picture of the essential regions influencing predictions across all three models, see in Figures 9, 10, and 11 . These visualizations serve a crucial role, not just in making our models transparent, but also in fostering a deeper understanding and instilling trust in how decisions are made.

6. Conclusions

In essence, the objective of this investigation was to assess the proficiency of diverse DL models in categorizing brain tumors. Following a series of comprehensive experiments and detailed analysis, it becomes apparent that these models exhibit varying degrees of competence for the assigned task. Models such as DenseNet121, EfficientNetB0, ResNet50 GoogleNet, and Inception V3 distinguished themselves as top performers with almost flawless levels of accuracy, precision, recall, and F1 scores. For detailed classification results, refer to Table 3. Conversely, AlexNet and the innovative ViT Transformer a recent contender in the field, displayed potential but fell behind in terms of accuracy and achieving an optimal equilibrium between precision and recall. This research accentuates the significance of carefully selecting the most suitable DL model that aligns with the specific requirements of the application. It further underscores how advancements in neural network architectures, exemplified by the ViT Transformer, persist in shaping the field of DL and computer vision, presenting captivating prospects for future advancements.

In summary, both Grad-Cam and Grad-CAM++ consistently provide a more acceptable insight into model interpretability compared to other methods tested in our study. Put simply, these methods precisely reveal the location of tumors, significantly enhancing our understanding of how DL models make decisions in classifying brain tumors. Therefore, it can be concluded that Grad-Cam and Grad-CAM++ heatmaps have improved our interpretative accuracy, playing a pivotal role in refining our understanding of DL model decision-making processes. These methodologies have been instrumental in enhancing the precision of our interpretations. This study contributes to selecting the correct DL model for brain tumor classification tasks while shedding light on ongoing challenges in making these models transparent and interpretable. Our contributions are outlined as follows:

- Model Evaluation: The study comprehensively assesses various DL architectures, providing valuable insights into which models are most effective for brain tumor classification. This evaluation is crucial for guiding the selection of appropriate models in real-world medical imaging applications.
- Brain Tumor Diagnosis: Diagnosing a brain tumor is a challenging process that requires the correct and rapid examination of MRI scan images. The study's findings directly contribute to enhancing the accuracy and reliability of DL models for identifying brain tumors, focusing on this specific medical area. This is critical for early diagnosis and treatment planning for patients.
- Model Interpretability: The incorporation of explainability approaches, such as Grad-Cam, Grad-Cam++, Integrated Gradient, and Saliency Mapping, represents a significant scientific contribution. By using these methods, the study increases the interpretability of DL models, shedding light on the decision-making processes and providing valuable intuition into how these models arrive at their classifications, particularly in the context of brain tumor diagnosis.

 doi:10.20944/preprints202402.0960.v1

Author Contributions: Conceptualization: W.N., M.A., and J.C.N.; methodology: W.N., and M.A.; software: W.N., M.A., and Y.B.; validation: M.A., Y.B., and J.C.N.; writing—original draft preparation: W.N., M.A., and J.C.N.; writing—review and editing: W.N., M.A., Y.B. and J.C.N.; supervision: M.A., and J.C.N.; funding: W.N. and J.C.N.

Data Availability Statement: Publicly available datasets were analyzed in this study. This data can be found at https://www.kaggle.com/datasets/sartajbhuvaji/brain-tumor-classification-mri. (accessed on 20 July 2023).

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

Abbreviation	Expansion
CNN	Convolutional Neural Network
CT	Computed Tomography
DL	Deep Learning
DeconvNET	Deconvolution network
DeepLIFT	Deep Learning Important FeaTures
F1 Score	Harmonic Precision-Recall Mean
Grad-Cam	Gradient-weighted Class Activation Mapping
GBP	Guided back propagation
LRP	Layer-wise relevance propagation
MRI	Magnetic Resonance Imaging
ReLU	Rectified Linear Unit
SHAP	SHapley Additive exPlanation
TL	Transfer Learning
VGG	Visual Geometry Group
XAI	Explainable Artificial Intelligence

References

1. D. Bau, B. Zhou, A. Khosla, A. Oliva, A. Torralba, Network dissection: Quantifying interpretability of deep visual representations, in: Proceedings of the IEEE conference on computer vision and pattern recognition,2017, pp. 6541–6549.
2. Abiwinanda, N., Hanif, M., Hesaputra, S.T., Handayani, A. and Mengko, T.R., 2019. Brain tumor classification using convolutional neural network. In World Congress on Medical Physics and Biomedical Engineering 2018: June 3-8, 2018, Prague, Czech Republic (Vol. 1) (pp. 183-189). Springer Singapore.
3. Ghazanfar Latif et al. "Multiclass brain Glioma tumor classification using block-based 3D Wavelet features of MR images." In: 2017 4th International Conference on Electrical and Electronic Engineering (ICEEE). IEEE. 2017, pp. 333–337.
4. Fukushima, K., 1980. Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. Biological cybernetics, 36(4), pp.193-202.
5. Brima, Y. and Atemkeng, M., 2022. Visual Interpretable and Explainable Deep Learning Models for Brain Tumor MRI and COVID-19 Chest X-ray Images. arXiv preprint arXiv:2208.00953.
6. Ebiele, J., Ansah-Narh, T., Djiokap, S., Proven-Adzri, E., & Atemkeng, M. (2020, September). Conventional machine learning based on feature engineering for detecting pneumonia from chest X-rays. In Conference of the South African Institute of Computer Scientists and Information Technologists 2020 (pp. 149-155).
7. Brima, Y., Atemkeng, M., Tankio Djiokap, S., Ebiele, J., & Tchakounté, F. (2021). Transfer learning for the detection and diagnosis of types of pneumonia including pneumonia induced by COVID-19 from chest X-ray images. Diagnostics, 11(8), 1480.
8. Chattopadhay, A.; Sarkar, A.; Howlader, P.; Balasubramanian, V.N. Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks. In Proceedings of the 2018 IEEE Winter Conference on Applications of Computer Vision (WACV), Lake Tahoe, NV, USA, 12–15 March 2018; pp. 839–847
9. 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. In Proceedings of the IEEE international conference on computer vision (pp. 618-626).

10. Sundararajan, M., Taly, A. and Yan, Q., 2017, July. Axiomatic attribution for deep networks. In International conference on machine learning (pp. 3319-3328). PMLR.
11. Zhang, X., Yao, J., Wang, Y., and Zhang, L. (2019). Brain Tumor Classification Using Deep Learning Neural Networks. Journal of Healthcare Engineering, 2019, 1-8.
12. Zhang, Z., Xie, Y., Xing, F., and McGough, M. (2021). Interpretability of deep learning models for medical image analysis: A survey. Medical Image Analysis, 101934.
13. Adebayo, J., Gilmer, J., Muelly, M., Goodfellow, I., Hardt, M., and Kim, B. (2018). Sanity Checks for Saliency Maps. In NeurIPS.
14. Kim, B., Wattenberg, M., Gilmer, J., Cai, C., Wexler, J., Viegas, F., and Sayres, R. (2020). Interpretability Beyond Feature Attribution: Quantitative Testing with Concept Activation Vectors (TCAV). In ICML.
15. Gao, J., Liang, J., Song, Z., and Xie, Y. (2019). Explainable AI in Radiology: Current Status and Future Directions. Bioengineering, 6(4), 94.
16. Cheplygina, V., de Bruijne, M., Pluim, J. P., and Marchiori, E. (2021). A Survey of Uncertainty Quantification in Deep Learning for Medical Image Analysis. Medical Image Analysis, 101880.
17. Zeiler, M.D. and Fergus, R., 2014. Visualizing and understanding convolutional networks. In Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part I 13 (pp. 818-833). Springer International Publishing.
18. Zhang, Q., Wang, W. and Zhu, S.C., 2018, April. Examining CNN representations with respect to dataset bias. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 32, No. 1).
19. Nguyen, A., Clune, J., Bengio, Y., Dosovitskiy, A. and Yosinski, J., 2017. Plug and play generative networks: Conditional iterative generation of images in latent space. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 4467-4477).
20. Fong and Vedaldi, 2017. Ruth C. Fong and Andrea Vedaldi. Interpretable explanations of black boxes by meaningful perturbation. In ICCV, 2017.
21. Zeineldin, RA, Karar, ME, Elshaer, Z, Coburger, ·J, Wirtz, CR, Burgert, O, et al. Explainability of deep neural networks for MRI analysis of brain tumors. Int J Comput Assist Radiol Surg. (2022) 17:1673–83. doi: 10.1007/s11548-022-02619-x.
22. Philbrick, KA, Yoshida, K, Inoue, D, Akkus, Z, Kline, TL, Weston, AD, et al. What does deep learning see? Insights from a classifier trained to predict contrast enhancement phase from CT images. AJR Am J Roentgenol. (2018) 211:1184–93. doi: 10.2214/ajr.18.20331.
23. Martí-Juan, G, Frías, M, Garcia-Vidal, A, Vidal-Jordana, A, Alberich, M, Calderon, W, et al. Detection of lesions in the optic nerve with magnetic resonance imaging using a 3D convolutional neural network. Neuroimage Clin. (2022) 36:103187. doi: 10.1016/j.nicl.2022.103187.
24. Zeiler, M., and Fergus, R. Visualizing and understanding convolutional networks. arXiv:1311.2901 (2013). doi: 10.48550/arXiv.1311.2901.
25. Chatterjee, S, das, A, Mandal, C, Mukhopadhyay, B, Vipinraj, M, Shukla, A, et al. TorchEsegeta: framework for interpretability and explainability of image-based deep learning models. Appl Sci. (2022) 12:2022. doi: 10.3390/app12041834.
26. Springenberg, J., Dosovitskiy, A., Brox, T., and Riedmiller, M. Striving for simplicity: the all convolutional net. arXiv:1412.6806 (2014). doi: 10.48550/arXiv.1412.6806.
27. Wood, DA, Kafiabadi, S, Busaidi, AA, Guilhem, E, Montvila, A, Lynch, J, et al. Deep learning models for triaging hospital head MRI examinations. Med Image Anal. (2022) 78:102391. doi: 10.1016/j.media.2022.102391.
28. Saleem, H, Shahid, AR, and Raza, B. Visual interpretability in 3D brain tumor segmentation network. Comput Biol Med. (2021) 133:104410. doi: 10.1016/j.compbiomed.2021.104410.
29. Montavon, G, Binder, A, Lapuschkin, S, Samek, W, and Müller, K-R. Layer-wise relevance propagation: an overview In: W Samek, G Montavon, A Vedaldi, L Hansen, and KR Müller, editors. Explainable AI: interpreting, explaining and visualizing deep learning. Lecture notes in computer science. Cham: Springer (2019). 193–209.
30. Golla, AK, Tönnes, C, Russ, T, Bauer, DF, Froelich, MF, Diehl, SJ, et al. Automated screening for abdominal aortic aneurysm in CT scans under clinical conditions using deep learning (2021) Diagnostics, 11:2131. doi: 10.3390/diagnostics11112131.

31. Shi, W, Tong, L, Zhu, Y, and Wang, MD. COVID-19 automatic diagnosis with radiographic imaging: explainable attention transfer deep neural networks. IEEE J Biomed Health Inform. (2021) 25:2376–87. doi: 10.1109/jbhi.2021.3074893.
32. Karim, MR, Jiao, J, Dohmen, T, Cochez, M, Beyan, O, Rebholz-Schuhmann, D, et al. DeepKneeExplainer: explainable knee osteoarthritis diagnosis from radiographs and magnetic resonance imaging. IEEE Access. (2021) 9:39757–80. doi: 10.1109/ACCESS.2021.3062493.
33. Lopatina, A, Ropele, S, Sibgatulin, R, Reichenbach, JR, and Güllmar, D. Investigation of deep-learning-driven identification of multiple sclerosis patients based on susceptibility-weighted images using relevance analysis. Front Neurosci. (2020) 14:609468. doi: 10.3389/fnins.2020.609468.
34. Shrikumar, A., Greenside, P., and Kundaje, A. Learning important features through propagating activation differences. arXiv:1704.02685 (2017). doi: 10.48550/arXiv.1804.02391.
35. Gulum, MA, Trombley, CM, and Kantardzic, M. A review of explainable deep learning cancer detection models in medical imaging. Appl Sci. (2021) 11:2021–5. doi: 10.3390/app11104573.
36. Singh, A, Sengupta, S, and Lakshminarayanan, V. Explainable deep learning models in medical image analysis. J Imaging. (2020) 6:52. doi: 10.3390/jimaging6060052.
37. Wang, C, Ma, J, Shao, J, Zhang, S, Liu, Z, Yu, Y, et al. Predicting EGFR and PD-L1 status in NSCLC patients using multitask AI system based on CT images. Front Immunol. (2022) 13:813072. doi: 10.3389/fimmu.2022.813072.
38. Kumar, A, Manikandan, R, Kose, U, Gupta, D, and Satapathy, SC. Doctor's dilemma: evaluating an explainable subtractive spatial lightweight convolutional neural network for brain tumor diagnosis. ACM Trans Multimedia Comput Commun Appl. (2021) 17:1–26. doi: 10.1145/3457187.
39. Uyulan, C, Erguzel, TT, Turk, O, Farhad, S, Metin, B, and Tarhan, N. A class activation map-based interpretable transfer learning model for automated detection of ADHD from fMRI data. Clin EEG Neurosci. (2022):15500594221122699. doi: 10.1177/15500594221122699.
40. Wang, CJ, Hamm, CA, Savic, LJ, Ferrante, M, Schobert, I, Schlachter, T, et al. Deep learning for liver tumor diagnosis part II: convolutional neural network interpretation using radiologic imaging features. Eur Radiol. (2019) 29:3348–57. doi: 10.1007/s00330-019-06214-8.
41. Akatsuka, J, Yamamoto, Y, Sekine, T, Numata, Y, Morikawa, H, Tsutsumi, K, et al. Illuminating clues of cancer buried in prostate MR image: deep learning and expert approaches. Biomolecules. (2019) 9:673. doi: 10.3390/biom9110673.
42. Fuhrman, JD, Gorre, N, Hu, Q, Li, H, El Naqa, I, and Giger, ML. A review of explainable and interpretable AI with applications in COVID-19 imaging. Med Phys. (2022) 49:1–14. doi: 10.1002/mp.15359.
43. Alshazly, H, Linse, C, Barth, E, and Martinetz, T. Explainable COVID-19 detection using chest CT scans and deep learning. Sensors. (2021) 21:455. doi: 10.3390/s21020455.
44. Hao, J, Xie, J, Liu, R, Hao, H, Ma, Y, Yan, K, et al. Automatic sequence-based network for lung diseases detection in chest CT. Front Oncol. (2021) 11:781798. doi: 10.3389/fonc.2021.781798.
45. Lahsaini, I, El Habib Daho, M, and Chikh, MA. Deep transfer learning based classification model for covid-19 using chest CT-scans. Pattern Recognit Lett. (2021) 152:122–8. doi: 10.1016/j.patrec.2021.08.035.
46. Garg, A, Salehi, S, Rocca, M, Garner, R, and Duncan, D. Efficient and visualizable convolutional neural networks for COVID-19 classification using chest CT. Expert Syst Appl. (2022) 195:116540. doi: 10.1016/j.eswa.2022.116540.
47. Ullah, F, Moon, J, Naeem, H, and Jabbar, S. Explainable artificial intelligence approach in combating real-time surveillance of COVID19 pandemic from CT scan and X-ray images using ensemble model. J Supercomput. (2022) 78:19246–71. doi: 10.1007/s11227-022-04631-z.
48. Lu, SY, Zhang, Z, Zhang, YD, and Wang, SH. CGENet: a deep graph model for COVID-19 detection based on chest CT. Biology. (2022) 11:2022–1. doi: 10.3390/biology11010033.
49. Jadhav, S, Deng, G, Zawin, M, and Kaufman, AE. COVID-view: diagnosis of COVID-19 using chest CT. IEEE Trans Vis Comput Graph. (2022) 28:227–37. doi: 10.1109/tvcg.2021.3114851.
50. Nagaoka, T, Kozuka, T, Yamada, T, Habe, H, Nemoto, M, Tada, M, et al. A deep learning system to diagnose COVID-19 pneumonia using masked lung CT images to avoid AI-generated COVID-19 diagnoses that include data outside the lungs. Adv Biomed Eng. (2022) 11:76–86. doi: 10.14326/abe.11.76.

51. Suri, JS, Agarwal, S, Chabert, GL, Carriero, A, Paschè, A, Danna, PSC, et al. COVLIAS 20-cXAI: cloud-based explainable deep learning system for COVID-19 lesion localization in computed tomography scans. Diagnostics. (2022) 12:1482. doi: 10.3390/diagnostics12061482.
52. Pennisi, M, Kavasidis, I, Spampinato, C, Schinina, V, Palazzo, S, Salanitri, FP, et al. An explainable AI system for automated COVID-19 assessment and lesion categorization from CT-scans. Artif Intell Med. (2021) 118:102114. doi: 10.1016/j.artmed.2021.102114.
53. Zhang, X, Han, L, Zhu, W, Sun, L, and Zhang, D. An explainable 3D residual self-attention deep neural network for joint atrophy 718 localization and Alzheimer's disease diagnosis using structural MRI. IEEE J Biomed Health Inform. (2021) 26:5289–97. doi: 719 10.1109/jbhi.2021.3066832.
54. Li, CF, Xu, YD, Ding, XH, Zhao, JJ, du, RQ, Wu, LZ, et al. MultiR-net: a novel joint learning network for COVID-19 segmentation and classification. Comput Biol Med. (2022) 144:105340. doi: 10.1016/j.compbiomed.2022.105340.
55. Williamson, BJ, Khandwala, V, Wang, D, Maloney, T, Sucharew, H, Horn, P, et al. Automated grading of enlarged perivascular spaces in clinical imaging data of an acute stroke cohort using an interpretable, 3D deep learning framework. Sci Rep. (2022) 12:788. doi: 10.1038/s41598-021-04287-4.
56. Kim, KH, Koo, HW, Lee, BJ, Yoon, SW, and Sohn, MJ. Cerebral hemorrhage detection and localization with medical imaging for cerebrovascular disease diagnosis and treatment using explainable deep learning. J Korean Phys Soc. (2021) 79:321–7. doi: 10.1007/s40042-021-00202-2.
57. Singh, A, Kwiecinski, J, Miller, RJH, Otaki, Y, Kavanagh, PB, van Kriekinge, S, et al. Deep learning for explainable estimation of mortality risk from myocardial positron emission tomography images. Circ Cardiovasc Imaging. (2022) 15:e014526. doi: 10.1161/circimaging.122.014526.
58. Jain, V, Nankar, O, Jerrish, DJ, Gite, S, Patil, S, and Kotecha, K. A novel AI-based system for detection and severity prediction of dementia using MRI. IEEE Access. (2021) 9:154324–46. doi: 10.1109/ACCESS.2021.3127394.
59. Hu, M, Qian, X, Liu, S, Koh, AJ, Sim, K, Jiang, X, et al. Structural and diffusion MRI based schizophrenia classification using 2D pretrained and 3D naive convolutional neural networks. Schizophr Res. (2022) 243:330–41. doi: 10.1016/j.schres.2021.06.011
60. Zhang, X, Han, L, Zhu, W, Sun, L, and Zhang, D. An explainable 3D residual self-attention deep neural network for joint atrophy localization and Alzheimer's disease diagnosis using structural MRI. IEEE J Biomed Health Inform. (2021) 26:5289–97. doi: 10.1109/jbhi.2021.3066832.
61. Simonyan, K., Vedaldi, A. and Zisserman, A., 2013. Deep inside convolutional networks: Visualising image classification models and saliency maps. arXiv preprint arXiv:1312.6034.
62. A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in Proc. Adv. Neural Inf. Process. Syst., 2012, pp.
63. He, K., Zhang, X., Ren, S. and Sun, J., 2016. Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 770-778).
64. Huang, G., Liu, Z., Van Der Maaten, L. and Weinberger, K.Q., 2017. Densely connected convolutional networks. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 4700-4708).
65. Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V. and Rabinovich, A., 2015. Going deeper with convolutions. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 1-9).
66. Chollet, F. Xception: Deep Learning with Depthwise Separable Convolutions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Washington, DC, USA, 19–24 June 2016; p. 1610-02357.
67. Szegedy, C.; Vanhoucke, V.; Ioffe, S.; Shlens, J.; Wojna, Z. Rethinking the inception architecture for computer vision. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Washington, DC, USA, 19–24 June 2016; pp. 2818–2826.
68. Simonyan, K. and Zisserman, A., 2014. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556.
69. Mascarenhas, S. and Agarwal, M., 2021, November. A comparison between VGG16, VGG19 and ResNet50 architecture frameworks for Image Classification. In 2021 International conference on disruptive technologies for multi-disciplinary research and applications (CENTCON) (Vol. 1, pp. 96-99). IEEE.

 doi:10.20944/preprints202402.0960.v1

70. Tan, M. and Le, Q., 2019, May. Efficientnet: Rethinking model scaling for convolutional neural networks. In International conference on machine learning (pp. 6105-6114). PMLR.
71. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S. and Uszkoreit, J., 2020. An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929.
72. Islam, MN, Hasan, M, Hossain, MK, Alam, MGR, Uddin, MZ, and Soylu, A. Vision transformer and explainable transfer learning models for auto detection of kidney cyst, stone and tumor from CT-radiography. Sci Rep. (2022) 12:11440. doi: 10.1038/s41598-022-15634-4.
73. www.kaggle.com/datasets/sartajbhuvaji/brain-tumor-classification-mri. (accessed on 20 May 2023)