npj | precision oncology

Review article

Published in partnership with The Hormel Institute, University of Minnesota

https://doi.org/10.1038/s41698-024-00789-2

A review of deep learning for brain tumor analysis in MRI

Check for updates

Felix J. Dorfner[1], Jay B. Patel[1], Jayashree Kalpathy-Cramer[2], Elizabeth R. Gerstner[1,3] & Christopher P. Bridge[1] ✉

Recent progress in deep learning (DL) is producing a new generation of tools across numerous clinical applications. Within the analysis of brain tumors in magnetic resonance imaging, DL finds applications in tumor segmentation, quantification, and classification. It facilitates objective and reproducible measurements crucial for diagnosis, treatment planning, and disease monitoring. Furthermore, it holds the potential to pave the way for personalized medicine through the prediction of tumor type, grade, genetic mutations, and patient survival outcomes. In this review, we explore the transformative potential of DL for brain tumor care and discuss existing applications, limitations, and future directions and opportunities.

Deep learning (DL), a form of artificial intelligence (AI), is rapidly transforming various fields, demonstrating remarkable success in tackling complex challenges, such as image recognition and natural language processing. These capabilities of DL have also found applications within medicine, with DL models having demonstrated effectiveness on tasks such as medical text summarization, prediction of future lung cancer risk, prediction of SARS-CoV-2 infectivity and variant evolution, identification of new antibiotics, and assessment of mammography for breast cancer[1–5]. In this review, we specifically focus on the applications of DL to brain tumor image analysis where there have been several important advances as well.

Brain tumors are the most common solid tumors in children and adolescents. Annually, more than 88,000 adults and 5500 children are diagnosed with brain tumors in the United States alone. These tumors have very high mortality, with a 5-year relative survival rate following diagnosis of a malignant brain or other CNS tumor of only 35.6% in adults[6].

According to the 5th edition of the WHO classification, tumors of the central nervous system (CNS) are classified into different tumor grades based on histological, immunohistochemical, and molecular features[7]. Diffuse gliomas (which include Astrocytomas, Oligodendrogliomas, and Ependymomas) are the most common type of primary malignant CNS tumor in adults, making up about 25% of all such cases[6]. The most common primary non-malignant brain tumors in adults are meningiomas. And the overall most common type of brain tumor is brain metastases, as about 20% of all patients with cancer will develop brain metastases during the course of their treatment[8].

Within neuro-oncology, the need for DL-powered solutions is directly related to the complexities associated with brain tumors. Brain tumors exhibit substantial heterogeneity in their presentation and require diverse therapeutic approaches. Analyzing these tumors accurately and efficiently is crucial for optimizing patient care. DL can play a pivotal role in this regard in one of two broad capacities. The first set of applications of DL models is on tasks that are very time-consuming for human experts within the existing clinical workflow, such as the creation of 3D segmentation masks for tumor quantification. However, beyond this, DL models have been shown to be capable of extracting insights beyond human capabilities, such as the prediction of important genomic biomarkers based on MRI alone[9]. As these capacities mature and develop, DL may help shape the workflows of tomorrow.

This review explores the transformative impact of DL on brain tumor analysis, focusing on its applications in two broad areas: segmentation and classification. We discuss how DL models are enabling automated and accurate tumor segmentation from medical images, facilitating objective and reproducible measurements crucial for diagnosis, treatment planning, and disease monitoring.

We also provide an outlook on current innovations for medical DL models. Namely, we discuss the growing role of foundation models, which are models trained on massive datasets of diverse data types, in the field of medicine. We anticipate that these models will greatly enhance the accuracy of DL-based brain tumor analysis and enable researchers to extend it toward tumor types for which analysis was previously unfeasible due to the limited amount of training data available.

Deep learning methods for MRI analysis

As a general term, AI refers to the development of computer systems capable of performing tasks that typically require human intelligence, such as visual perception, decision-making, and learning.

[1]Athinoula A. Martinos Center for Biomedical Imaging, 149 13th St, Charlestown, MA, 02129, USA. [2]University of Colorado School of Medicine, Anschutz Medical Campus, Aurora, CO, 80045, USA. [3]Massachusetts General Hospital Cancer Center, Boston, MA, 02114, USA. ✉e-mail: cbridge@mgh.harvard.edu

The Hormel Institute
University of Minnesota

The first generation of AI applications for MRI data included *radiomics*-based approaches. Here, predefined sets of features—quantifying image characteristics such as intensity, contrast, shape or texture—were extracted from the image and then passed into a classical machine learning model, such as random forests or support vector machines, to make predictions[10,11]. Their application to brain tumors has been extensively covered by previous reviews[12]. By contrast, DL models use artificial neural networks to learn complex patterns and relationships *directly* from the data, during a process called *training*. During training, a model iteratively improves by adjusting the internal model parameters to minimize the difference between its predictions and the known ground-truth labels. By analyzing vast amounts of information, these deep learning models can identify subtle features and make predictions that may not be readily apparent to human observers. Since features in DL models are optimized directly for a particular task, they are often able to make better predictions than the previous generations of radiomics-based approaches.

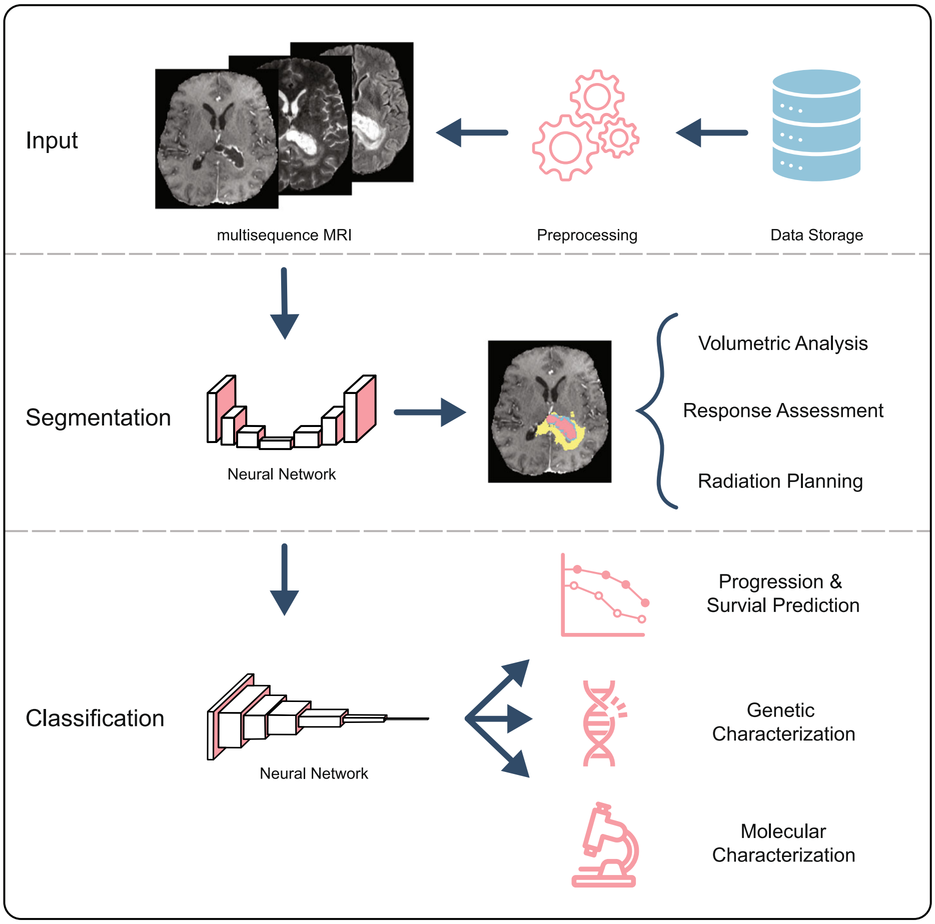
The majority of modern deep learning-based architectures for image analysis are based on convolutional neural networks (CNN)[13]. The core component of a convolutional neural network is a convolutional layer, which moves a learned filter across the image in a "sliding-window" fashion. This approach uses fewer parameters compared to a simple "fully-connected" neural network and is able to recognize the learned patterns regardless of their location in the image. Each convolutional layer in the network allows for increasing abstraction and identification of more complex features. Within the scope of image analysis, tasks can be broadly split into two groups: image segmentation (i.e. delineating salient/relevant regions on the image) and image classification (i.e. categorizing the image from a set of pre-determined classes). As illustrated in Figure 1, these deep learning models have shown state-of-the-art performance across a wide range of neuro-oncology tasks, including but not limited to tumor segmentation, prediction of mutation status for gliomas, and prediction of treatment response[14–16].

Since brain MRIs are volumetric images, it is natural to employ convolutions along all three dimensions, giving rise to 3D CNNs[17]. While this is the more common approach for analyzing brain MRI and allows the model to fully consider volumetric information when making predictions, there are some drawbacks compared to the 2D CNNs that are more common in other application domains. Firstly, 3D networks typically have more parameters, leading to higher computational demands and making them more challenging to optimize especially on small datasets. Additionally, powerful 2D networks pre-trained on large datasets of 2D images are readily available to use as the basis of new models via transfer learning[18], but a limited choice of 3D architectures is available in this way. Finally, since thick slice MRI data is highly anisotropic, it is not necessarily optimal to treat all three spatial dimensions equivalently within the model. Thus some authors use 2D CNNs applied to each 3D slice of the MRI, potentially with immediately adjacent slices included as additional channels in a so-called "2.5D" architecture. Mixed results suggest that the optimal choice of architecture may be dependent on application and training dataset[19–21].

Segmentation and quantification

Perhaps the most well-studied application of DL within neuro-oncology is that of tumor segmentation. Segmentation is the process of delineating tumor regions (or sub-regions) within an image and is a key step in tumor quantification, response assessment, and treatment planning, as well as a preliminary step for further analyses of different tumor regions (see the section "Classification").

Fig. 1 | Examples of deep-learning-based workflows for MRI segmentation and classification. For the segmentation task, the CNN receives an input image, often consisting of multiple sequences, and outputs a segmentation map according to the given task, such as segmenting a tumor. For the classification task, the model receives the input image and outputs a classification into two or more classes.



Public datasets and benchmarks

The proliferation of work on segmentation has been made possible thanks to the wide availability of relevant publicly accessible data. Publicly available datasets allow researchers around the world to train DL models on multi-institutional, high-quality datasets and thus serve as a benchmark for comparing models. The largest public datasets of brain tumor MRI images are listed in Tables 1–3.

Some of the most widely used public datasets come from the Brain Tumor Segmentation (BraTS) challenge, hosted annually by the Medical Image Computing and Computer-Assisted Intervention (MICCAI) Society since 2012. Initially focused solely on the segmentation of gliomas, BraTS has expanded over the years to include other CNS tumors. In 2022, BraTS introduced a pediatric dataset and a dataset of adult-type diffuse glioma of underrepresented patients (BraTS-Africa)[22,23], and in 2023, challenges for the segmentation of brain metastases and meningiomas[22,24] were added. Due to the scope, size, and accessibility of these datasets, BraTS has become an important benchmark for state-of-the-art brain tumor segmentation.

One challenge is that public datasets can greatly vary in quality and content. For example, consider datasets available for brain metastases segmentation projects. While some provide just a single imaging sequence (e.g. MOLAB[25]), others may provide multiple sequences (e.g. UCSF-BMSR[26]). Some may provide the raw images (e.g. Brain-TR-GammaKnife[27]), whereas others may provide only pre-processed versions (e.g. NYUMets[28]). Differences in annotations may also exist, with some datasets providing binarized (tumor vs. no tumor) labels (e.g. BrainMetShare[29]) and others providing multi-class (contrast-enhancing tumor, necrosis, and peritumoral edema) labels (e.g. BraTS-METS[30]). As such, every dataset is composed of images and annotations tailored towards a specific endpoint, which can create non-trivial problems for combining datasets from different sources to solve a single task such as segmentation.

Segmentation approaches and architectures

As the BraTS challenge is an important benchmark for the performance of segmentation models, examining the winning architectures of the past challenges provides an overview of the evolution of medical image segmentation architectures.

One fundamental segmentation model is the U-Net, which was first introduced by Ronneberger et al. in 2015[31] and has been the basis of the winning segmentation models in BraTS since then as well as becoming ubiquitous in other medical image segmentation tasks such as segmenting intracranial metastases[32] and many other biological structures[33]. The U-Net is composed of two main structures: the contracting path gradually downsamples the image and extracts features at lower spatial resolution and a higher semantic level, and then the expanding path re-combines these features to create the segmentation mask by gradually upsampling the image again to the input resolution. Skip connections between layers in the contracting path and the expanding path are used to preserve detailed information lost during downsampling, ensuring fine detail in the output segmentation mask.

Various modifications to the standard U-Net have since been proposed to improve performance. A crucial early improvement was the introduction of the Dice loss by Milletari et al.[17]. The Dice loss function, based on the similar Dice similarity coefficient (DSC) measures the overlap between the predicted foreground region and the ground truth foreground region regardless of the size of the foreground region. As a result, it is suitable for segmentation problems in which there exists a large class imbalance between the foreground and background classes, and/or in which foreground sizes vary considerably between different samples in the dataset, unlike the common cross-entropy loss. In 2018, Myronenko placed first in the BraTS segmentation challenge by utilizing an asymmetrical U-Net with residual blocks, which contain shortcuts within the network that help preserve information and improve learning during training[34]. In 2019, the winning architecture was a cascade of two similar U-Nets, where the additional second model was used to refine the coarse segmentation maps generated from the first[35]. In 2020, Isensee et al. proposed No-New-Net (nnU-Net), which built a framework around a standard U-Net architecture and automated most of the deep learning pipeline including image preprocessing, model adaptation, hyperparameter tuning, and an ensembling strategy, leading to improved performance and consistency[33]. Ensembling is

Table 1 | Overview of public datasets for MRI studies of brain tumors

Public Dataset	Data Publisher	No. of Cases/Patients	Tumor Type
BraTS 2021[a93]	RSNA-ASNR-MICCAI	2000 patients	Adult diffuse glioma
BraTS-Africa 2023[23]	MICCAI-CAMERA- Lacuna Fund	95 cases	Adult diffuse glioma
BraTS-PEDs 2023[22]	CBTN-CONNECT- DIPGR-ASNR-MICCAI	228 patients	Pediatric high-grade glioma
BraTS Meningioma 2024[24]	RSNA-ASNR-MICCAI	1650 cases	Meningioma
BraTS-METS 2023[b30]	RSNA-ASNR-MICCAI	328 cases	Brain metastasis
NYUMets[28]	New York University	1429 patients	Brain metastasis
UPenn-GBM[132]	University of Pennsylvania	630 patients	Glioblastoma
UCSF-BMSR[26]	University of California San Francisco	412 patients	Brain metastasis
TCGA-GBM[133]	TCGA Glioma Phenotype Research Group	262 patients	Glioblastoma
Figshare Dataset[134]	Southern Medical University, Guangzhou, China	233 patients	Glioma, meningioma, pituitary
GLIS-RT Dataset[82]	Massachusetts General Hospital	230 patients	Glioblastoma, astrocytoma, low-grade glioma
Pretreat-MetsToBrain-Masks[135]	Yale School of Medicine	200 patients	Brain metastasis
TCGA-LGG[136]	TCGA Glioma Phenotype Research Group	199 patients	Low-grade glioma
BrainMetShare[29]	Stanford University	156 patients	Brain metastasis
CPTAC-GBM[137]	National Cancer Institute's Clinical Proteomic Tumor Analysis Consortium	99 patients	Glioblastoma
MOLAB[25]	University of Castilla-La Mancha	75 patients	Brain metastasis
Brain-TR-Gamma-Knife[27]	University of Mississippi	47 patients	Brain metastasis
IVY GAP[138]	Allen Institute for Brain Science	39 patients	Glioblastoma

[a] Contains a subset of studies from the TCGA-GBM, TCGA-LGG, IVY GAP, and CPTAC-GBM Datasets.

[b]Subset of studies from the NYUMets, UCSF-BMSR, Pretreat-MetsToBrain-Masks, and BrainMetsShare Datasets.

Table 2 | Overview of model architectures, training data, and metrics results from selected papers

Rerence	Model architecture name	Training data used	Test set results (Dice)
Myronenko et al.[34]	Asymmetrical U-Net	BraTS 2018	WT: 88.39, TC: 81.54, ET: 76.64
Jiang et al.[35]	Two-Stage Cascaded U-Net	BraTS 2019	WT: 88.80, TC: 83.70, ET: 83.27
Isensee et al.[33]	nnU-Net (no new-Net)	BraTS 2020	WT: 88.95, TC: 85.06, ET: 82.03
Luu and Park[36]	modified nnU-Net	BraTS 2021	WT: 92.75, TC: 87.81, ET: 84.51
Zeineldin et al.[14]	Ensemble: DeepSeg, nnU-Net, and DeepSCAN	BraTS 2022	WT: 92.94, TC: 87.88, ET: 88.03

WT whole tumor, *TC* tumor core, *ET* enhancing tumor.

Table 3 | Overview of model architectures, training data, and metrics results from selected papers for classification tasks

Reference	Dataset	Input	No. Patients	Task	Metric (result)
Recurrence vs. radiation necrosis					
Gao et al.[87]	Private	T1, T1-c, T2	146	Recurrence vs. radiation necrosis	AUC: 0.915
Lee et al. (2020)[88]	Private	T1, T1-c, T2, T2-FSE, FLAIR, ADC	46	Recurrence vs. radiation necrosis	AUC: 0.81
Survival prediction					
McKinley et al. (2020)[89]	BraTS 2020	T1, T1-c, T2, FLAIR	587	Overall survival (>15, 15–10, <10 months)	Accuracy: 0.617
Yan et al. (2023)[90]	BraTS 2019	T1, T1-c, T2, FLAIR	205	Overall survival (>15, 15–10, <10 months)	Accuracy: 0.548
IDH status prediction					
Chang et al. (2018)[9]	Private/TCIA	T1, T1-c, T2, FLAIR	291	IDH status prediction	AUC: 0.95
Choi et al. (2020)[15]	Private/TCIA	T1-c, T2, FLAIR	1166	IDH status prediction	AUC: 0.86–0.96
Molecular biomarker prediction					
Tak et al. (2024)[91]	Private/CBTN	T2	326	BRAF mutational status prediction	AUC: 0.73–0.82
Calabrese et al. (2020)[92]	Private	T1, T1-c, T2, FLAIR	199	9 Molecular biomarkers prediction	AUC: 0.55–0.97
Chen et al. (2020)[94]	TCIA	T1-c/FLAIR	106	MGMT methylation status prediction	AUC: 0.828–0.897
Yogananda et al. (2021)[95]	TCIA	T2	247	MGMT methylation status prediction	AUC: 0.93
Chen et al. (2022)[96]	Private	T1, T1-c, T2, ADC	111	MGMT methylation status prediction	AUC: 0.9
Saeed et al. (2022)[97]	BraTS 2021	T1, T1-c, T2, FLAIR	585	MGMT methylation status prediction	AUC: 0.54–0.64
Robinet et al. (2023)[98]	BraTS 2021/Private	T1, T1-c, T2, FLAIR	672	MGMT methylation status prediction	AUC: 0.60–0.65

a technique in which different models are trained to perform the same task and their individual outputs are combined into one final prediction. The segmentation winner in 2021 improved on this nnU-Net solution by incorporating an asymmetric contracting path to achieve a better balance of resources between a more powerful contracting path and a simpler expanding path, group normalization for more robust training with the small batch sizes necessary when using large 3D images, and an axial attention decoder focus to model attention on relevant parts of the input image[36]. This improved architecture was ensembled together with DeepSeg and DeepScan to create an even stronger solution for the 2022 segmentation challenge[14,37,38].

Object detection networks, which output coarse bounding box predictions rather than detailed voxel-wise classification, may be used as an alternative to segmentation. This may be more appropriate where the detection of lesions is more important than quantification and measurement, and has the key advantage that they can be trained using simple bounding box annotations, which are much quicker to perform than segmentation masks. For example, Zhang et al.[39] use the Faster R-CNN architecture[40] for the detection of brain metastases. Furthermore, in some approaches such as DeSeg[41], object detection outputs can be used to screen for initial locations that can be fed to further models to perform detailed segmentation.

Multi-sequences and missing sequences

Typically, CNNs for MRI image analysis are trained on a specific sequence or set of sequences. Similar to a trained radiologist, the model processes different sequences of a study to make its decision, as these sequences highlight different aspects of the tumor. However, this multi-sequence training can be a limitation at inference time if an imaging study does not have all the required sequences. Therefore, several approaches have been developed to deal with missing sequences for MRI imaging studies in order to increase the applicability of trained models.

One approach includes network architectures that are designed to accommodate variable input sequences and have been explicitly trained for this task. One example of this is the Hetero-modal Image Segmentation approach by Havaei et al.[42]. Here, the model processes each sequence input independently to create a high-level representation of each image. These representations are then combined via simple operators (such as the mean or standard deviation) in such a way as to ensure that an arbitrary number of sequences can be provided to the model at test time. The combined representation is then passed to further layers of the network for segmentation. Using a different approach, Feng et al. directly trained a model to handle missing data by randomly replacing input sequences with empty images. This allowed the model to adapt to missing sequences and still perform the segmentation if one or more required sequences were missing[43].

A second approach generates missing sequences so that they can then be used in models that expect a fixed set of input sequences. State-of-the-art methods mostly use generative adversarial networks (GAN) to generate the missing target sequence from the available input sequences. Two important differences between sequence generation models are which sequences are required to generate the missing sequence and whether the input images need to be spatially aligned. One approach used this idea to generate T1-c and Double inversion recovery (DIR) images from three common input sequences (T1, T2, FLAIR)[44]. In addition to GANs, diffusion-based models, which gradually remove noise from a random image to generate realistic data, have also been proposed for the generation of missing MRI sequences[45].

Site generalization

It has been shown that DL model performance can suffer significantly if applied to data from different sites, and MRI is particularly vulnerable to this issue due to its flexibility and differences across scanners and protocols[46]. One approach to address this is harmonization, where techniques like StarGAN and CycleGAN standardize image appearance across sites to improve model robustness[47–50]. StarGAN and CycleGAN are both generative adversarial networks (GAN) that are trained by having the image generation model ("generator") compete with a model that tries to identify which images are real and which are synthetic ("discriminator"). As training in this way progresses, the generator generates increasingly realistic images. Another strategy is to apply extensive augmentation during training, adding diverse artifacts to expose the model to a range of imaging conditions. Methods such as SynthSeg, SynthMorph, and SynthStrip demonstrate that this approach can improve generalization across sites[51–53] for multiple tasks.

Pediatric brain tumors

Pediatric brain tumors represent the most common cause of cancer-related mortality in children[22]. Although some parallels exist with adult brain tumors, pediatric tumors often exhibit distinct imaging characteristics and clinical presentations. For instance, adult glioblastomas (GBMs) and pediatric DMGs are both high-grade gliomas associated with poor prognoses; however, their incidence rates and typical locations differ. GBMs, with an incidence of ~3 per 100,000, predominantly affect older adults and are frequently found in the frontal and/or temporal lobes, whereas DMGs are considerably rarer and typically arise in the pons of children aged 5–10 years. Furthermore, characteristic imaging features like post-gadolinium enhancement and necrosis, common in GBMs, are less consistently observed in DMGs, particularly at initial diagnosis[54,55]. Consequently, specialized imaging tools are critical for characterizing and assessing these pediatric tumors, deep learning models that were developed adult brain tumors can not just be applied to pediatric brain tumors, due to the differences between tumor presentations. As a result, separate deep-learning models need to be developed for most applications of segmenting and analyzing pediatric brain tumors. The BRATS challenge included a dataset of pediatric tumors for the first time in 2023. This dataset contains MRI sequences (T1, T1-c, T2, FLAIR) for 228 patients with pediatric high-grade gliomas. The winning team here achieved a mean dice score of 0.65 for ET, 0.81 TC, and 0.83 for segmenting the whole tumor[56].

Segmentation evaluation metrics

There are a variety of methods and metrics for evaluating the performance of deep-learning models. Choosing the appropriate metric for a given problem is crucial to ensure that it accurately captures the clinically relevant aspects of the task. A recent study proposed a framework called Metrics Reloaded, which provides a tool to guide researchers through the process of choosing the right validation metrics for their DL model[57].

The two most important metrics for brain tumor segmentation are the Dice score and the Hausdorff distance (HD). The Dice score measures the overlap between the predicted segmentation and the ground truth segmentation. It ranges from 0 (no overlap) to 1 (perfect overlap) and provides an intuitive assessment of segmentation accuracy that measures only the degree of overlap between the prediction and ground truth, regardless of the absolute size of the region. The Hausdorff distance, on the other hand, quantifies the maximum distance between any point on the predicted segmentation boundary and the nearest point on the ground truth boundary. As a distance metric, the HD ranges from 0 to infinity. To mitigate the effect of outliers, the 95th percentile HD is often used. While these metrics capture the agreement between the model prediction and the ground truth, it has been shown that the scores may have a poor correlation with clinician perception of segmentation quality[58]. Furthermore, currently, DL-based segmentation is often corrected by clinicians to ensure quality. Here, clinicians may prefer certain patterns in the model's segmentation behavior, for example over-segmentation rather than under-segmentation, as it may be easier to correct. These practical preferences are not captured by popular metrics of segmentation performance but may play an important role in clinical adoption[59].

$$\text{Dice Similarity Coefficient (DSC)} = \frac{2\,|GT \cap MS|}{|GT| + |MS|} \tag{1}$$

$$GT = \text{Ground Truth Surface}$$

$$MS = \text{Masked Segmentation}$$

$$\text{Average Hausdorff Distance (AHD)} = \max\left(d\left(S_{\mathrm{GT}}, S_{\mathrm{MS}}\right),\, d\left(S_{\mathrm{MS}}, S_{\mathrm{GT}}\right)\right) \tag{2}$$

$$S_{\mathrm{GT}} = \text{Ground Truth Surface}$$

$$S_{\mathrm{MS}} = \text{Masked Segmentation Surface}$$

$$d = \text{Distance Function}$$

$$\text{Accuracy} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}} \tag{3}$$

$$TP = \text{True Positives}$$

$$TN = \text{True Negatives}$$

$$FP = \text{False Positives}$$

$$FN = \text{False Negatives}$$

$$\text{AUROC} = \int_0^1 \mathrm{TPR}(t)\, d\mathrm{FPR}(t) \tag{4}$$

$$TPR = \text{True Positive Rate}$$

$$FPR = \text{False Positive Rate}$$

Uncertainty

Another important aspect of the clinical application of deep-learning-based segmentation models is the quantification of segmentation uncertainty. This approach can help clinicians when they are manually revising model segmentations, as areas marked as uncertain are most likely to contain mistakes. The implementation of uncertainty can also help to build trust in DL models. As part of the BraTS challenge in 2020, a metric to compare the performance of uncertainty measures was introduced, and a variety of solutions to the problem were submitted by participants[60].

McKinley et al. proposed a novel loss function (a function that measures the difference between the model prediction and the ground-truth) for the task of uncertainty estimation in brain tumor segmentation[38]. In their approach, the model outputted two probabilities for each voxel in the input image; one that the predicted label was correctly identifying the ground truth label and one that the predicted label did not correspond to the ground-truth label. These two outputs were used during model training to jointly optimize tumor segmentation and uncertainty estimation. Other groups have directly utilized the scores that the model outputs for each possible label in the segmentation mask to derive uncertainty scores[61,62]. These types of approaches can be further improved by combining models into an ensemble, leveraging the strengths of each individual model for a more accurate and robust outcome.

Other methods of quantifying uncertainty are based on the idea of creating a distribution of outputs for a single input image and using statistics of that distribution to estimate uncertainty. In Monte Carlo dropout, a fraction of the nodes in the network are randomly deactivated ("dropped out") and inference is performed a number of times, producing a distribution of results[63]. The variance in this distribution can be used to quantify the uncertainty in the prediction. Zhou et al. incorporated this technique into a 3D U-Net to obtain an uncertainty map for brain tumor segmentation. This uncertainty was then used as an additional input in a second stage to improve segmentation performance on the BraTS 2018 and 2019 dataset[64]. In test time augmentation, input images are transformed through random transforms, such as flipping, rotation, and scaling. The inference is then performed for each of the transformed views of the original input. This creates a distribution of outputs for each input image. Statistics about this distribution, such as variance end entropy can then be used to estimate uncertainty[65]. This approach has also been used for quantifying the uncertainty of brain tumor segmentation[66].

Generalizability plays an important role in medical DL application, as trained models may be applied to datasets with different characteristics, such as disease manifestation or image acquisition techniques. This change in the dataset is commonly referred to as domain shift. Hoebel et al. demonstrated the stability of uncertainty estimates for brain tumor segmentation quality assessment even under domain shift between high- and low-grade gliomas[67]. This suggests the potential for reliable uncertainty-based quality control tools in clinical practice, although further investigation is needed to confirm generalizability across various scenarios.

Quantification

Quantification of tumor segmentations plays an important role in the clinical applications of DL-based segmentation models. Metrics such as tumor dimensions and volume are regularly used as criteria for diagnostic and disease monitoring purposes. Examples of this include the response assessment in neuro-oncology (RANO) and the response assessment in pediatric neuro-oncology (RAPNO)[68–70]. This score is derived from 2D measurements of the tumor's maximal diameter and is used to assess treatment response in brain tumors. An automated deep-learning approach using a U-Net-based segmentation model showed high correlation with human raters for both 2D and volumetric segmentations[71,72].

One important challenge with manual dimension measurements is their poor inter-rater reproducibility[73]. This can potentially be solved by employing automated deep-learning-based measurements, that have shown superior repeatability and reliability compared to human readers[72]. Furthermore it has been shown that manual 2D RANO measurements are inferior compared to 3D volumetric measurements[16]. DL-based segmentation models play an important role in the clinical application of these volumetric evaluations as they provide high-quality annotations many orders of magnitude faster than a human annotator. Indeed, deep-learning-based assessment of tumor response has been shown to be a significantly better predictor of overall survival compared to RANO assessment[16].

Response assessment

Comparing the most recent study to prior patient imaging is a core component of a radiologist's workflow for brain tumor assessment. It has been argued that most current DL algorithms are not suitable for applications in tasks where comparisons with previous images are necessary[74]. Developing DL algorithms that can process longitudinal imaging data, therefore, plays an important role towards advancing DL in radiology.

Kickingereder et al. developed an application-ready software infrastructure for deep-learning-based segmentation of brain tumors[16]. They utilized spatial and temporal tumor volume dynamics to predict patient time to progression. This approach included the functionality to track lesions across time points and consider new lesions. This tracking is necessary for longitudinal volume and progression monitoring based on individual lesions. Oermann et al. developed an architecture that can segment brain metastases using longitudinal imaging data[28]. The architecture, which they call "segmentation through time", uses a collection of U-Nets. For time points after the baseline, the segmentation network also incorporates information about previous time points propagated by convolutional long short-term memory (LSTM) blocks. LSTM blocks are a type of neural network architecture that can learn and remember patterns over long sequences of data, making them well-suited for tasks involving time-series data, such as analyzing changes in medical images over multiple time points. Patel et al. developed a joint image registration and segmentation network called SPIRS to segment a new time point scan using prior time point information[75]. In their network, prior time point imaging is affinely and deformably warped onto the new time point image. The warped prior time point annotation is then used as a coarse initialization for the segmentation of the new time point. This approach significantly improves the segmentation performance for micro-metastatic brain lesions.

Another approach for estimating the treatment response of a tumor over time is through tumor growth modeling. Cell proliferation and migration within tumors can be mathematically modeled using a set of partial differential equations. However, their clinical application is hindered by the challenge of accurately estimating model parameters and other factors such as the initial tumor cell density from medical images. Recent research has explored DL-based approaches to overcome these limitations and improve parameter estimation efficiency and accuracy[76,77]. This has made it possible to construct models that are able to obtain accurate representations of tumor cell distribution and proliferation parameters from a single MRI scan. Predicting tumor invasion has important consequences for radiotherapy planning, as it would enable radiation oncologists to more accurately define radiation margins around the tumor, potentially targeting more of the tumor while reducing damage to healthy tissue[77].

Radiation therapy

Together with surgical resection and chemotherapy, radiation therapy is a fundamental component of treatment planning for many brain tumors. In order to maximize the effect on the tumor while minimizing adverse side effects of the radiation, the tumor and *organs at risk*, such as the brainstem, eyes and optic chiasm need to be carefully outlined on imaging in order to optimize the dose distribution. Additionally so called *barrier structures*, anatomical structures that are natural barriers to tumor spread (i.e. falx cerebri) are delineated in the imaging to optimize the margin around the actual tumor in which radiation is applied. While radiation planning is primarily performed on CT, MRI may also be acquired and used for enhanced soft-tissue contrast to aid in the delineation of the target and surrounding structures. In clinical practice these processes can be quite time-consuming, with human annotators requiring about 20 min to perform the contouring of relevant structures for a single patient[78]. Thus, there exists a clinical need for fast, human-level delineation through automation. Different deep-leaning-based solutions have been proposed for this purpose[78–81] and the GLIS-RT open dataset is available for benchmarking model development[82]. DL-based methods have achieved excellent results in segmenting larger structures, such as tentorium cerebelli, brain sinuses, or ventricles, with clinically acceptable accuracy. Furthermore DL models have been shown to be more consistent in their outputs compared to human experts[81]. However further development is needed to provide clinically acceptable segmentations of smaller structures such as the optic apparatus[78].

A particular challenge with radiation planning segmentation is the presence of post-treatment changes, such as resection cavities, blood products, and gliosis[83], which are under-represented in existing datasets and segmentation models, with a recent survey finding that over 98% of published research on glioma segmentation using pre-surgical imaging[58]. The 2024 BraTS challenge, for the first time, focuses on the important task of segmentation of post-treatment brain MRIs[83], and as such research, this area is likely to receive considerable attention in the near future.

Once the target, organs at risk and barrier structures are delineated, deep learning methods may additionally assist in selecting parameters for generating the radiation treatment plan, however to our knowledge this has only been demonstrated with CT imaging[84,85].

Classification

Beyond segmentation, applications of DL to brain tumor imaging can also include classification tasks. The most common architecture in medical image classification tasks is the ResNet, introduced by He et al. in 2015[86]. It first uses convolutional filters to extract important features from the image, similar to how the U-Net does for segmentation. These features are then passed through a series of layers, each containing multiple convolutional filters that identify increasingly complex patterns. Shortcut connections within these layers help preserve information and improve learning. Finally, a classification layer analyzes these extracted features and assigns the image to one of the predefined categories.

Since much of the development of DL models relies upon the availability of data, there has been less work in this area compared to segmentation, as there are fewer publicly available datasets.

Distinguishing tumor recurrence and radiation necrosis

Distinguishing tumor recurrence from radiation necrosis presents a significant challenge in glioma management. This distinction is clinically critical as it entails fundamentally different treatment approaches. However diagnosing this accurately can be difficult even for experienced clinicians as both conditions can exhibit similar features on conventional MRI. This challenge arises from the complex and often subtle differences in tumor appearance and tissue response to radiation. In contrast, deep learning models excel at analyzing intricate patterns in imaging data. Studies have developed different CNNs that use multi-modal MRI data to distinguish between recurrence and radiation necrosis[87,88]. They show promising results, with one model significantly surpassing the accuracy of clinicians on this task[87].

Survival prediction

Another area in which DL promises to enhance the clinical management of brain tumors is survival prediction. There are two popular DL-based approaches to the task: a multi-class classification problem or a Cox proportional hazards model. From 2017 to 2020, the BraTS challenge included a task focused on predicting the overall survival (in days) of glioma patients who had undergone gross tumor resection from MRI data. The top-performing approach from 2020 employed a two-stage strategy. First, a segmentation model was used to delineate the tumor and its sub-compartments. Next, features derived from the number of disconnected tumor segments, along with the patient's age, were inputted into both a linear regression model and a random forest classifier. These models were then combined, achieving an accuracy of 60% on the test dataset for classifying the patients into long-term survivors (>15 months), mid-term survivors (10–15 months), and short-term survivors (<10 months)[89]. A different study explored the use of a convolutional denoising autoencoder (DAE) network combined with a Cox proportional hazards model for survival prediction in glioblastoma patients. The DAE was used to extract features from multi-modal MRI data, which were then fed into the Cox model for survival analysis. This approach achieved a concordance index (C-index) of 0.74 on the test set[90]. Although these findings from both approaches demonstrate the potential of DL for survival prediction, achieving the level of accuracy required for robust clinical implementation remains an open challenge.

Biomarker prediction

One of the most promising applications of DL in brain tumor analysis is the prediction of genetic biomarkers directly from imaging data. This idea carries transformative potential for clinical practice, as it could enable clinicians to obtain important information about a tumor's genetic profile without the need for invasive biopsies or surgeries.

One example is the prediction of isocitrate dehydrogenase (IDH) mutation status in gliomas. IDH mutation status is an important factor in determining the prognosis and treatment of gliomas, and being able to identify it pre-treatment can significantly impact clinical decision-making. Deep learning-based methods have shown promise in predicting IDH status from MRI scans, offering a non-invasive alternative to traditional biopsy-based methods[9,15]. A model that first performed automated tumor segmentation and subsequently used both radiomics and deep-learning derived features was able to predict the IDH mutational status of patients diagnosed with gliomas with an accuracy of 78.8% and 93.8% on internal and external test sets, respectively. This model used three different MRI sequences (T1 post-contrast, T2 and FLAIR) as its input[15]. A recent publication explored the ability of a deep learning model to predict the mutational status of the BRAF gene in patients with pediatric low-grade gliomas based on T2-weighted MRI scans. The model was able to classify BRAF status into three classes (BRAF fusion, BRAF V600E, and wild-type) with an accuracy of 75% and 77% on internal and external test sets, respectively[91].

Calabrese et al. implemented an approach in which the tumor sub-compartments (i.e. enhancing tumor, non-enhancing tumor, and edema) were first segmented by a deep-learning model and subsequently used as the basis for radiomics feature extraction. These features were then passed to a random forest regression model. The authors explored the predictive capabilities of the model for nine different genetic biomarkers in glioblastoma. The model showed good results for predicting IDH mutations, ATRX mutations, chromosome 7/10 aneuploidies, and CDKN2 family mutations. The sensitivity for those biomarkers ranged from 0.76 to 0.94 and the specificity from 0.86 to 0.92[92].

A widely discussed use-case is the prediction of O6-methylguanine-DNA methyltransferase (MGMT) promoter methylation status, which is a key indicator of response to temozolomide chemotherapy in glioblastoma. There is a large public dataset created for the BraTS 2021 challenge, which provides information on the MGMT promoter methylation status along with MRI scans for 2040 patients[93]. However, the feasibility of accurately predicting MGMT status from MRI data remains controversial. While some studies have reported promising results[94–96], others have questioned the validity of these findings and argued that predicting MGMT status from MRI alone may not be possible with current techniques[97,98]. This highlights the importance of critically examining results and ensuring transparency in DL research, especially when considering clinical applications. Notably, in the 2021 BraTS challenge, the winning model for MGMT prediction achieved an AUROC (area under the receiver operating characteristic curve) of only 0.62, which is considered poor and certainly not sufficient for reliable clinical decision-making[99]. Such examples underscore the need for rigorous validation and cautious interpretation of DL models before integrating them into clinical workflows. Overall, the use of DL to predict genetic biomarkers from imaging data is a rapidly evolving field with significant potential to improve the diagnosis and treatment of brain tumors.

Future directions

While DL has demonstrated remarkable progress in brain tumor analysis, the field continues to evolve rapidly, with several promising avenues for future development.

Quantitative MRI

Quantitative MRI methods[100,101], while not widely deployed clinically, hold promise to increase the standardization of images between vendors by directly quantifying tissue properties, though other qualitative variations are likely to persist. Further, early evidence suggests that such images may provide further insight into tumor characteristics, such as infiltration beyond the contrast-enhancing region visible in conventional qualitative images[102]. However, currently, there is a lack of experimental studies demonstrating the value of quantitative MRI for developing AI models. For example, Tampu et al.[103] found no statistically significant advantage of utilizing quantitative relaxometry images over conventional T1 and T2 weighted images for developing AI models for the detection and identification of brain tumor biomarkers, however, the study was conducted on a very small dataset of 23 patients. Future work should focus on investigating the value of quantitative MRI for AI model development across larger datasets and multiple vendors.

Multimodal integration

The radiographic appearance of a tumor in an MR image cannot capture the full complexity of a brain tumor in its full clinical context. Consideration of other clinical information, including patient demographics, genomics, and histopathology is therefore crucial for clinical decision-making. However, despite the fact that deep learning models can naturally integrate many high-dimensional data types, most current work considers only a single modality. DL has already shown promise in analyzing other data types from brain tumor patients. For instance, deep learning models trained on whole-slide images (WSI) of histopathology slides have achieved high accuracy in predicting 1p/19q codeletion status in gliomas, surpassing the performance of traditional methods like fluorescence in situ hybridization (FISH)[104]. Initial studies have explored the integration of WSI data with genomic and transcriptomic information to predict survival outcomes in glioma patients[105], but there remains considerable potential for improved predictions by leveraging multi-modal models.

Vision transformers

A recent and important development is the transformer architecture[106]. While initially developed for applications related to natural language processing, it was adapted towards image-based tasks by Dosovitskiy et al.[107]. Due to the more flexible design of their proposed vision transformer (ViT) architecture, it is better able to capture long-range interactions within an image, meaning it can understand relationships between distant parts of the image, which is important for tasks like identifying complex shapes or patterns. As such, ViTs have been shown to outperform CNNs when given sufficiently large amounts of training data[107]. Since most brain tumor datasets are small, the potential benefits are yet to be realized. However, as the availability of large dataset sizes improves, ViTs may become increasingly used for brain tumor image analysis.

Foundation models

Another promising future direction is the development of foundation models for brain tumor imaging. Foundation models are an emerging paradigm in DL that involves the self-supervised training of large, general-purpose models on very large datasets of diverse data types[108]. These models learn fundamental patterns and relationships within the data, enabling them to perform a wide range of downstream tasks with remarkable accuracy and efficiency. Unlike traditional DL models that are trained for specific tasks, foundation models are adaptable and can be fine-tuned for different applications without requiring extensive retraining. Chen et al. recently introduced a general-purpose foundation model for pathology that was pretrained on more than 100 million images acquired from over 100,000 diagnostic H&E-stained WSIs across 20 different tissue types[109]. This model, after fine-tuning with limited task-specific data, achieved excellent performance on a range of tasks, including brain metastasis detection, glioma IDH1 mutation prediction, and histomolecular subtyping. The model showed excellent performance even on few-shot tasks for which only between 1 and 32 task-specific training examples per class were provided to the model. This is orders of magnitude fewer examples than would be needed without a foundation model. Another recent publication proposed a foundation model for cancer imaging biomarker discovery using computed tomography (CT) data from over 11,000 radiographic lesions[110]. After fine-tuning with limited data, their model outperformed other state-of-the-art models on a variety of tasks such as predicting malignancy in lung nodules and predicting survival in non-small cell lung cancer (NSCLC). Although their analysis did not include brain lesions or MRI images, the results demonstrate the potential of similar foundation models for MRI in neuro-oncology applications where large datasets are not available. The adoption of such foundation models is likely to accelerate future research on brain tumor analysis.

Limitations and challenges

Despite the significant progress over recent years in the application of DL to brain tumor analysis, only a small number of brain tumor-related models are approved for clinical use within the United States[111]. There remain substantial challenges to further research progress and its translation.

Datasets

It remains challenging to collate imaging datasets for applications beyond those covered by the existing public datasets[112]. Though vitally important, patient privacy and consent are the most important barriers to the widespread sharing of medical imaging data outside of individual hospitals and radiology providers, which individually see relatively small numbers of patients compared to those needed to train accurate deep learning models. In addition to the images themselves, curating or creating appropriate and accurate ground truth presents a further challenge, as the process is typically time-consuming and requires considerable expertise, with 3D segmentations, in particular, being slow to generate. Where manual processes are involved, there is often considerable inter-reader variability, which may or may not be clinically significant depending on context. Previous studies have found substantial variation between readers in obtaining quantitative measurements manually from MRIs[113–115]. Even when large datasets are created and released publicly, the utility of those images is limited by the availability of accompanying ground truth. For example a dataset of segmented brain MRIs containing tumors cannot be used for a study on outcomes prediction if no information on outcomes was released.

As a consequence, most studies in brain tumor analysis either focus on one of the existing public datasets, leading to the over-representation of three associated tasks in the literature, or use small, often single-institutional datasets that lack diversity. In particular, the availability of large datasets has led to a focus on gliomas, brain metastases, and meningiomas at the expense of work relating to rarer brain tumors, such as craniopharyngiomas, pineoblastomas, or parameningeal rhabdomyosarcomas.

Major initiatives such as the US National Cancer Institute's Imaging Data Commons (IDC)[116] and the European Federation for Cancer Images (EUCAIM), as well as challenges, such as BraTS will continue to serve an important role in collating and standardizing access to imaging data at scale for the research community. Furthermore, federated learning[117,118], a technique wherein models are trained across multiple sites without data leaving each site, will likely play an increasing role in model development in order to create large datasets while retaining patient privacy but it is not without its own technological and logistical challenges.

Reproducibility of research

The high variability of brain MRI data coupled with the small number of cases often used in (frequently private) datasets leads to this area being particularly vulnerable to issues surrounding lack of reproducibility. In many cases, published articles have not yet been replicated by further studies, and in some cases further studies have been conducted but failed to replicate the previous findings[97]. Results obtained on small, private datasets, should be treated with caution as they fail to generalize beyond a single institution or MRI scanner and may be the result of unintended bias or spurious correlation present in the training data (for example, between a clinical outcome and the scanner used to acquire the image), or may be the result of chance. In some cases methodological failures may have given rise to reporting of incorrect results: as an example, a paper on the prediction of MGMT status from gliomas was withdrawn upon the discovery of an error in the computer code used to conduct the study[119], and a review of machine learning in radiomic analyses (not using DL) identified common mistakes that inflate performance[120].

Where possible, the public release of datasets and source code can help to reduce the likelihood of replication issues in research.

Domain shift and generalizability

As noted above, DL models can fail to generalize beyond the sites and scanners represented within their training data, which creates a considerable challenge for deploying DL at scale, and the flexibility and complexity of MRI as a modality make this a particular concern for brain tumor analysis[46]. In addition to changes between sites, changes over time at a single site are due to factors including scanner software, imaging protocols, clinical workflows, and patient demographics, and this, in turn, can lead to performance degradation[121,122].

Bias and fairness

Another crucial consideration is fairness. While there is no single definition of fairness within AI, generally speaking, it refers to unequal model performance on different subpopulations, for example, different races, genders, or ages[123,124]. Unequal model performance can, in turn, lead to unequal outcomes in clinical care. The causes of unfairness may simply be under-representation of subpopulations within the training data, but it can also be more insidious and difficult to avoid. Biases and outcome differences that exist within healthcare—and therefore in model training data —can be propagated by artificial intelligence models[125,126], because models can learn to infer protected characteristics, such as age, race, and gender even if they not provided directly to a DL model[127], and then learn to associate these with, for example, poorer outcomes that are a result of socio-economic factors.

Therefore, model developers, especially those developing models intended for clinical use, should, therefore, follow established guidelines to screen their models for fairness across any relevant subpopulations[123,124].

Clinical translation

Although deep learning models have significant potential to benefit patient care, incorrect predictions or inappropriate use of DL models pose a significant risk of harm. Several steps are crucial to minimize harm[128], including thorough validation of models on representative data for clinical effectiveness and education of physicians in the capabilities and limitations of the technology[129] to reduce the risk of automation bias, where physicians blindly follow the predictions of an algorithm[130].

Explainability

Where DL models are to be used for cancer treatment decision-making, it is vital that their predictions are understandable to clinicians. Unfortunately, most DL models, including most articles in this review, provide black-box predictions. Building interpretable models remains one of the major unsolved technical challenges within the field of DL. Many approaches to the explainability of DL methods rely on determining which regions of the image are relevant to the prediction using techniques, such as saliency or occlusion maps. However, this level of explanation is likely to be insufficient for most of the applications discussed above. Counterfactual explanations, which allow the user to visualize how the image would need to change to change the prediction, may provide one more promising direction[131].

Received: 23 April 2024; Accepted: 17 December 2024;
Published online: 03 January 2025

References

1. Van Veen, D. et al. Adapted large language models can outperform medical experts in clinical text summarization. *Nat. Med.* 1–9 https://www.nature.com/articles/s41591-024-02855-5 (2024).
2. Mikhael, P. G. et al. Sybil: a validated deep learning model to predict future lung cancer risk from a single low-dose chest computed tomography. *J. Clin. Oncol.* **41**, 2191–2200 (2023).
3. Wang, G. et al. Deep-learning-enabled protein–protein interaction analysis for prediction of SARS-CoV-2 infectivity and variant evolution. *Nat. Med.* **29**, 2007–2018 (2023).
4. Wong, F. & Collins J. J. 'Explainable' AI identifies a new class of antibiotics. *Nature* https://www.nature.com/articles/d41586-023-03668-1 (2023).
5. McKinney, S. M. et al. International evaluation of an AI system for breast cancer screening. *Nature* **577**, 89–94 (2020).
6. Ostrom, Q. T., Cioffi, G., Waite, K., Kruchko, C. & Barnholtz-Sloan, J. S. CBTRUS Statistical Report: primary brain and other central nervous system tumors diagnosed in the United States in 2014–2018. *Neuro-Oncology* **23**, iii1–iii105 (2021).
7. Louis, D. N. et al. The 2021 WHO Classification of Tumors of the Central Nervous System: a summary. *Neuro-Oncology* **23**, 1231–1251 (2021).
8. Achrol, A. S. et al. Brain metastases. *Nat. Rev. Disease Primers* **5**, 1–26 (2019).
9. Chang, K. et al. Residual convolutional neural network for determination of IDH status in low- and high-grade gliomas from MR imaging. *Clin. Cancer Res.* **24**, 1073–1081 (2018).
10. Breiman, L. Random forests. *Mach. Learn.* **45**, 5–32 (2001).
11. Cortes, C. & Vapnik, V. Support-vector networks. *Mach. Learn.* **20**, 273–297 (1995).
12. Patel, J., Gidwani, M., Chang, K. & Kalpathy-Cramer, J. Radiomics and radiogenomics with deep learning in neuro-oncology. In Kia, S. M. et al. (eds) *Machine Learning in Clinical Neuroimaging and Radiogenomics in Neuro-oncology* Vol. 12449, 199–211 (Springer International Publishing, Cham, 2020).
13. Krizhevsky, A., Sutskever, I. & Hinton, G. E. ImageNet classification with deep convolutional neural networks. *Commun. ACM* **60**, 84–90 (2017).
14. Zeineldin, R.A., Karar, M.E., Burgert, O., Mathis-Ullrich, F. (2023). Multimodal CNN Networks for Brain Tumor Segmentation in MRI: A BraTS 2022 Challenge Solution. In *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries BrainLes 2022* (eds Bakas, S.) Vol. 13769 *of Lecture Notes in Computer Science* (Springer, Cham 2022).
15. Choi, Y. S. et al. Fully automated hybrid approach to predict the IDH mutation status of gliomas via deep learning and radiomics. *Neuro-Oncology* **23**, 304–313 (2020).
16. Kickingereder, P. et al. Automated quantitative tumour response assessment of MRI in neuro-oncology with artificial neural networks: a multicentre, retrospective study. *Lancet Oncol.* **20**, 728–740 (2019).
17. Milletari, F., Navab, N. & Ahmadi, S.-A. V-Net: fully convolutional neural networks for volumetric medical image segmentation. In *2016 Fourth International Conference on 3D Vision (3DV)* 565–571 (IEEE, Stanford, CA, USA, 2016).
18. Zhuang, F. et al. A comprehensive survey on transfer learning. *Proc. IEEE* **109**, 43–76 (2021).
19. Zhang, Y., Liao, Q., Ding, L. & Zhang, J. Bridging 2D and 3D segmentation networks for computation-efficient volumetric medical image segmentation: an empirical study of 2.5D solutions. *Comput. Med. Imaging Graphics* **99**, 102088 (2022).
20. Ottesen, J. A. et al. 2.5D and 3D segmentation of brain metastases with deep learning on multinational MRI data. *Front. Neuroinform.* **16**, 1056068 (2023).
21. Henschel, L. et al. FastSurfer—a fast and accurate deep learning based neuroimaging pipeline. *NeuroImage* **219**, 117012 (2020).
22. Kazerooni, A. F. et al. The Brain Tumor Segmentation (BraTS) challenge 2023: focus on pediatrics (CBTN-CONNECT-DIPGR-ASNR-MICCAI BraTS-PEDs) http://arxiv.org/abs/2305.17033 (2023).
23. Adewole, M. et al. The Brain Tumor Segmentation (BraTS) challenge 2023: glioma segmentation in sub-Saharan Africa patient population (BraTS-Africa) http://arxiv.org/abs/2305.19369 (2023).
24. LaBella, D. et al. The ASNR-MICCAI brain tumor segmentation (BraTS) challenge 2023: intracranial meningioma http://arxiv.org/abs/2305.07642 (2023).
25. Ocaña-Tienda, B. et al. A comprehensive dataset of annotated brain metastasis MR images with clinical and radiomic data. *Sci. Data* **10**, 208 (2023).
26. Rudie, J. D. et al. The University of California San Francisco Brain Metastases Stereotactic Radiosurgery (UCSF-BMSR) MRI Dataset. *Radiol. Artif. Intell.* **6**, e230126 (2024).
27. Wang, Y. et al. A brain MRI dataset and baseline evaluations for tumor recurrence prediction after Gamma Knife radiotherapy. *Sci. Data* **10**, 785 (2023).
28. Oermann, E. et al. *Longitudinal Deep Neural Networks for Assessing Metastatic Brain Cancer on a Massive Open Benchmark* https://www.researchsquare.com/article/rs-2444113/v1 (2023).

29. Grøvik, E. et al. Deep learning enables automatic detection and segmentation of brain metastases on multi-sequence MRI. *J. Magn. Reson. Imaging* **51**, 175–182 (2020).
30. Moawad, A. W. et al. The brain tumor segmentation (BraTS-METS) challenge 2023: brain metastasis segmentation on pre-treatment MRI. ArXiv:2306.00838 [eess, q-bio] version: 1. (2023).
31. Ronneberger, O., Fischer, P. Brox, T. U-Net: Convolutional Networks for Biomedical Image Segmentation. In *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015* (eds Navab, N., Hornegger, J., Wells, W., and Frangi, A.), Vol. 9351, *Lecture Notes in Computer Science* (Springer, Cham, 2015).
32. Rudie, J. D. et al. Three-dimensional U-Net convolutional neural network for detection and segmentation of intracranial metastases. *Radiology: Artif. Intell.* **3**, e200204 (2021).
33. Isensee, F., Jaeger, P. F., Full, P. M., Vollmuth, P. & Maier-Hein, K. H. nnU-Net for Brain Tumor Segmentation. In *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries BrainLes 2020* (eds Crimi, A., and Bakas, S.), Vol. 12659, *Lecture Notes in Computer Science* (Springer, Cham, 2021).
34. Myronenko, A. 3D MRI Brain Tumor Segmentation Using Autoencoder Regularization. In *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries BrainLes 2018* (eds Crimi, A., Bakas, S., Kuijf, H., Keyvan, F., Reyes, M., van Walsum, T.), Vol. 11384, *Lecture Notes in Computer Science* (Springer, Cham, 2019).
35. Jiang, Z., Ding, C., Liu, M. & Tao, D. Two-stage cascaded U-Net: 1st Place solution to BraTS challenge 2019 segmentation task. In Crimi, A. & Bakas, S. (eds) *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries*, Vol. 11992, *Lecture Notes in Computer Science* 231–241 (Springer International Publishing, Cham, 2020).
36. Luu, H.M. & Park, S. H. Extending nn-UNet for Brain Tumor Segmentation. In *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries BrainLes 2021* (eds Crimi, A., Bakas, S.), Vol. 12963, *Lecture Notes in Computer Science* (Springer, Cham., 2022).
37. Zeineldin, R. A., Karar, M. E., Mathis-Ullrich, F. & Burgert, O. Ensemble CNN networks for GBM tumors segmentation using multi-parametric MRI. In Crimi, A. & Bakas, S. (eds) *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries*, Vol. 12962, *Lecture Notes in Computer Science* 473–483 (Springer International Publishing, Cham, 2022).
38. McKinley, R., Meier, R. & Wiest, R. Ensembles of Densely-Connected CNNs with Label-Uncertainty for Brain Tumor Segmentation. In Crimi, A. et al. (eds.) *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries*, Vol. 11384, *Lecture Notes in Computer Science* 456–465 (Springer International Publishing, Cham, 2019).
39. Zhang, M. et al. Deep learning detection of cancer metastases to the brain on MRI. *J. Magn. Reson. Imaging* **52**, 1227–1236 (2020).
40. Ren, S., He, K., Girshick, R. & Sun, J. Faster R-CNN: towards real-time object detection with region proposal networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **39**, 1137–1149 (2017).
41. Yu, H. et al. DeSeg: auto detector-based segmentation for brain metastases. *Phys. Med. Biol.* **68**, 025002 (2023).
42. Havaei, M., Guizard, N., Chapados, N. & Bengio, Y. HeMIS: Hetero-Modal Image Segmentation. In *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2016*, (eds Ourselin, S., Joskowicz, L., Sabuncu, M., Unal, G. & Wells, W.) Vol. 9901, *Lecture Notes in Computer Science* (Springer, Cham, 2016).
43. Feng, X. et al. Brain tumor segmentation for multi-modal MRI with missing information. *J. Digit. Imaging* **36**, 2075–2087 (2023).
44. Li, H. et al. DiamondGAN: Unified Multi-modal Generative Adversarial Networks for MRI Sequences Synthesis. *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2019. Lecture Notes in Computer Science* Vol. 11767, (Springer, Cham, 2019).
45. Kim, J. & Park, H. Adaptive Latent Diffusion Model for 3D Medical Image to Image Translation: Multi-modal Magnetic Resonance Imaging Study. *2024 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 7589–7598 (Waikoloa, HI, USA, 2024), https://doi.org/10.1109/WACV57701.2024.00743.
46. Guo, B. et al. The impact of scanner domain shift on deep learning performance in medical imaging: an experimental study. https://arxiv.org/abs/2409.04368. Version Number: 2. (2024).
47. Komandur, D. et al. Unsupervised harmonization of brain MRI using 3D CycleGANs and its effect on brain age prediction. *2023 19th International Symposium on Medical Information Processing and Analysis (SIPAIM).* 1–5 (Mexico City, Mexico. 2023) https://doi.org/10.1109/SIPAIM56729.2023.10373501.
48. Roca, V. et al. IGUANe: A 3D generalizable CycleGAN for multicenter harmonization of brain MR images. *Med. Image Anal.* **99**, 103388 (2025).
49. Roca, V. et al. A three-dimensional deep learning model for inter-site harmonization of structural MR images of the brain: extensive validation with a multicenter dataset. *Heliyon* **9**, e22647 (2023).
50. Gebre, R. K. et al. Cross-scanner harmonization methods for structural MRI may need further work: a comparison study. *NeuroImage* **269**, 119912 (2023).
51. Billot, B. et al. SynthSeg: segmentation of brain MRI scans of any contrast and resolution without retraining. *Med. Image Anal.* **86**, 102789 (2023).
52. Hoffmann, M. et al. SynthMorph: learning contrast-invariant registration without acquired images. *IEEE Trans. Med. Imaging* **41**, 543–558 (2022).
53. Hoopes, A., Mora, J. S., Dalca, A. V., Fischl, B. & Hoffmann, M. SynthStrip: skull-stripping for any brain image. *NeuroImage* **260**, 119474 (2022).
54. Kazerooni, A. F. et al. Automated tumor segmentation and brain tissue extraction from multiparametric MRI of pediatric brain tumors: A multi-institutional study. *Neuro-Oncol. Adv.* **5**, vdad027 (2023).
55. Nabavizadeh, A. et al. Current state of pediatric neuro-oncology imaging, challenges and future directions. *Neoplasia* **37**, 100886 (2023).
56. Kazerooni, A. F. et al. BraTS-PEDs: results of the multi-consortium international pediatric brain tumor segmentation challenge 2023. http://arxiv.org/abs/2407.08855 (2024).
57. Maier-Hein, L. et al. Metrics reloaded: recommendations for image analysis validation. *Nat. Methods* **21**, 195–212 (2024).
58. Hoebel, K. V. et al. Expert-centered evaluation of deep learning algorithms for brain tumor segmentation. *Radiology: Artif. Intell.* **6**, e220231 (2024).
59. Hoebel, K. V. et al. Not without context—a multiple methods study on evaluation and correction of automated brain tumor segmentations by experts. *Acad. Radiol.* **31**, 1572–1582 (2024).
60. Mehta, R. et al. QU-BraTS: MICCAI BraTS 2020 challenge on quantifying uncertainty in brain tumor segmentation—analysis of ranking scores and benchmarking results. *Mach. Learn. Biomed. Imaging* **1**, 1–54 (2022).
61. Dai, C. et al. Self-training for brain tumour segmentation with uncertainty estimation and biophysics-guided survival prediction. In Crimi, A. & Bakas, S. (eds) *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries*, *Lecture Notes in Computer Science* 514–523 (Springer International Publishing, Cham, 2021).
62. Patel, J. et al. Segmentation, survival prediction, and uncertainty estimation of gliomas from multimodal 3D MRI using selective kernel networks. In Crimi, A. & Bakas, S. (eds) *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries*. *Lecture Notes in Computer Science* 228–240 (Springer International Publishing, Cham, 2021).
63. Gal, Y. & Ghahramani, Z. Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning. *Proceedings of*

The 33rd International Conference on Machine Learning, Proceedings of Machine Learning Research **48**, 1050–1059 (2016).

64. Zhou, T. & Zhu, S. Uncertainty quantification and attention-aware fusion guided multi-modal MR brain tumor segmentation. *Comput. Biol. Med.* **163**, 107142 (2023).
65. Wang, G. et al. Aleatoric uncertainty estimation with test-time augmentation for medical image segmentation with convolutional neural networks. *Neurocomputing* **338**, 34–45 (2019).
66. Wang, G. et al. Automatic Brain Tumor Segmentation Using Convolutional Neural Networks with Test-Time Augmentation. *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries, Lecture Notes in Computer Science***11384**, 61–72 (2019).
67. Hoebel, K. V. et al. Do I know this? segmentation uncertainty under domain shift. In Išgum, I. & Colliot, O. (eds) *Medical Imaging 2022: Image Processing* Vol. 27 (SPIE, San Diego, USA, 2022).
68. Wen, P. Y. et al. RANO 2.0: update to the response assessment in neuro-oncology criteria for high- and low-grade gliomas in adults. *J. Clin. Oncol.* **41**, 5187–5199 (2023).
69. Erker, C. et al. Response assessment in paediatric high-grade glioma: recommendations from the Response Assessment in Pediatric Neuro-Oncology (RAPNO) working group. *Lancet Oncol.* **21**, e317–e329 (2020).
70. Fangusaro, J. et al. Response assessment in paediatric low-grade glioma: recommendations from the response assessment in pediatric neuro-oncology (RAPNO) working group. *Lancet Oncol.* **21**, e305–e316 (2020).
71. Chang, K. et al. Automatic assessment of glioma burden: a deep learning algorithm for fully automated volumetric and bidimensional measurement. *Neuro-Oncology* **21**, 1412–1422 (2019).
72. Peng, J. et al. Deep learning-based automatic tumor burden assessment of pediatric high-grade gliomas, medulloblastomas, and other leptomeningeal seeding tumors. *Neuro-Oncology* **24**, 289–299 (2022).
73. Raman, F. et al. Evaluation of RANO criteria for the assessment of tumor progression for lower-grade gliomas. *Cancers* **15**, 3274 (2023).
74. Acosta, J. N., Falcone, G. J. & Rajpurkar, P. The need for medical artificial intelligence that incorporates prior images. *Radiology* **304**, 283–288 (2022).
75. Patel, J. et al. A Deep Learning based framework for joint image registration and segmentation of brain metastases on magnetic resonance imaging. In *Proc. 8th Machine Learning for Healthcare Conference*, Vol. 219 (eds Deshpande, K., Fitera, M., Joshi, S., Lipton, Z., Ranganath, R., Urteaga, I. & Yeung, S.) 565–587 *Proceedings of Machine Learning Research* (2023).
76. Ezhov, I. et al. Neural Parameters Estimation for Brain Tumor Growth Modeling. *Medical Image Computing and Computer Assisted Intervention – MICCAI 2019. Lecture Notes in Computer Science* **11765**, 787–795 (2019).
77. Martens, C. et al. Deep learning for reaction-diffusion glioma growth modeling: towards a fully personalized model? *Cancers* **14**, 2530 (2022).
78. Turcas, A. et al. Deep-learning magnetic resonance imaging-based automatic segmentation for organs-at-risk in the brain: accuracy and impact on dose distribution. *Phys. Imaging Radiat. Oncol.* **27**, 100454 (2023).
79. Agn, M. et al. A modality-adaptive method for segmenting brain tumors and organs-at-risk in radiation therapy planning. *Med. Image Anal.* **54**, 220–237 (2019).
80. Shusharina, N., Heinrich, M. P., Huang, R. In *Proc. Segmentation, Classification, and Registration of Multi-modality Medical Imaging Data: MICCAI 2020 Challenges, ABCs 2020, L2R 2020, TN-SCUI 2020, Held in Conjunction with MICCAI 2020*, Lima, Peru, October 4–8, 2020*,* Vol. 12587, *Lecture Notes in Computer Science* (Springer International Publishing, Cham, 2021).
81. Shusharina, N. et al. Cross-modality brain structures image segmentation for the radiotherapy target definition and plan optimization. In *Segmentation, Classification, and Registration of Multi-modality Medical Imaging Data*, Vol. 12587 (eds Shusharina, N., Heinrich, M. P. & Huang, R.) 3–15 (Springer International Publishing, Cham, 2021).
82. Shusharina, N. & Bortfeld, T. *Glioma Image Segmentation for Radiotherapy: RT Targets, Barriers to Cancer Spread, and Organs at Risk (GLIS-RT)* https://www.cancerimagingarchive.net/collection/glis-rt/ (2021).
83. Verdier, M. C. d. et al. The 2024 Brain Tumor Segmentation (BraTS) challenge: glioma segmentation on post-treatment MRI. ArXiv:2405.18368 [cs] (2024).
84. Tsang, D. S. et al. A pilot study of machine-learning based automated planning for primary brain tumours. *Radiat. Oncol.* **17**, 3 (2022).
85. Tsang, D. S. et al. A prospective study of machine learning-assisted radiation therapy planning for patients receiving 54 Gy to the brain. *Int. J. Radiat. Oncol.*Biol.*Phys.* **119**, 1429–1436 (2024).
86. He, K., Zhang, X., Ren, S. & Sun, J. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* 770–778 (eds Agapito, L., Berg, T., Kosecka, J. & Zelnik-Manor, L.) (IEEE, 2016).
87. Gao, Y. et al. Deep learning methodology for differentiating glioma recurrence from radiation necrosis using multimodal magnetic resonance imaging: algorithm development and validation. *JMIR Med. Inform.* **8**, e19805 (2020).
88. Lee, J. et al. Discriminating pseudoprogression and true progression in diffuse infiltrating glioma using multi-parametric MRI data through deep learning. *Sci. Rep.* **10**, 20331 (2020).
89. McKinley, R. et al. Uncertainty-Driven Refinement of Tumor-Core Segmentation Using 3D-to-2D Networks with Label Uncertainty. *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries. Lecture Notes in Computer Science* **12658**, 401–411 (2021).
90. Yan, T. et al. Survival prediction for patients with glioblastoma multiforme using a Cox proportional hazards denoising autoencoder network. *Front. Comput. Neurosci.* **16**, 916511 (2023).
91. Tak, D. et al. Noninvasive molecular subtyping of pediatric low-grade glioma with self-supervised transfer learning. *Radiology: Artif. Intell.* https://pubs.rsna.org/doi/10.1148/ryai.230333 (2024).
92. Calabrese, E., Villanueva-Meyer, J. E. & Cha, S. A fully automated artificial intelligence method for non-invasive, imaging-based identification of genetic alterations in glioblastomas. *Sci. Rep.* **10**, 11852 (2020).
93. Baid, U. et al. The RSNA-ASNR-MICCAI BraTS 2021 benchmark on brain tumor segmentation and radiogenomic classification. https://arxiv.org/abs/2107.02314 (2021).
94. Chen, X. et al. Automatic prediction of MGMT status in glioblastoma via deep learning-based MR image analysis. *BioMed Res. Int.* **2020**, 1–9 (2020).
95. Yogananda, C. G. B. et al. MRI-based deep-learning method for determining glioma MGMT promoter methylation status. *Am. J. Neuroradiol.* **42**, 845–852 (2021).
96. Chen, S. et al. Predicting MGMT promoter methylation in diffuse gliomas using deep learning with radiomics. *J. Clin. Med.* **11**, 3445 (2022).
97. Saeed, N., Hardan, S., Abutalip, K. & Yaqub, M. Is it possible to predict MGMT promoter methylation from brain tumor MRI scans using deep learning models? In *Proc. 5th International Conference on Medical Imaging with Deep Learning* Vol. 172 1005–1018 (eds Konukoglu, E., Menze, B., Venkataraman, A., Baumgartner, C., Dou, Q. & Albarqouni, S.) *Proceedings of Machine Learning Research* (2022).
98. Robinet, L., Siegfried, A., Roques, M., Berjaoui, A. & Cohen-Jonathan Moyal, E. MRI-based deep learning tools for MGMT promoter methylation detection: a thorough evaluation. *Cancers* **15**, 2253 (2023).

99. Flanders, A. et al. *RSNA-MICCAI Brain Tumor Radiogenomic Classification* https://kaggle.com/competitions/rsna-miccai-brain-tumor-radiogenomic-classification (2021).
100. Mills, A. F., Sakai, O., Anderson, S. W. & Jara, H. Principles of quantitative MR imaging with illustrated review of applicable modular pulse diagrams. *RadioGraphics* **37**, 2083–2105 (2017).
101. Gulani, V. & Seiberlich, N. Quantitative MRI: rationale and challenges. In *Advances in Magnetic Resonance Technology and Applications*, Vol. 1 (eds Seiberlich, N., Gulani, V., Calamante, F., Campbell-Washburn, A., Doneva, M., Hu, H. H. & Sourbron, S.) xxxvii–li (Elsevier, 2020).
102. Blystad, I. et al. Quantitative MRI for analysis of peritumoral edema in malignant gliomas. *PLoS ONE* **12**, e0177135 (2017).
103. Tampu, I. E., Haj-Hosseini, N., Blystad, I. & Eklund, A. Deep learning-based detection and identification of brain tumor biomarkers in quantitative MR-images. *Mach. Learn. Sci. Technol.* **4**, 035038 (2023).
104. Kim, G. J., Lee, T., Ahn, S., Uh, Y. & Kim, S. H. Efficient diagnosis of IDH-mutant gliomas: 1p/19qNET assesses 1p/19q codeletion status using weakly-supervised learning. *npj Precis. Oncol.* **7**, 1–9 (2023).
105. Steyaert, S. et al. Multimodal deep learning to predict prognosis in adult and pediatric brain tumors. *Commun. Med.* **3**, 1–15 (2023).
106. Vaswani, A. et al. Attention is all you need. In *Advances in Neural Information Processing Systems* Vol. 30 (eds von Luxburg, U., Guyon, I., Bengio, S., Wallach, H. & Fergus, R.) (Curran Associates, Inc., 2017).
107. Dosovitskiy, A. et al. An image is worth 16×16 words: transformers for image recognition at scale. https://arxiv.org/abs/2010.11929 (2021).
108. Moor, M. et al. Foundation models for generalist medical artificial intelligence. *Nature* **616**, 259–265 (2023).
109. Chen, R. J. et al. Towards a general-purpose foundation model for computational pathology. *Nat. Med.* **30**, 850–862 (2024).
110. Pai, S. et al. Foundation model for cancer imaging biomarkers. *Nat. Mach. Intell.* 1–14 https://www.nature.com/articles/s42256-024-00807-9 (2024).
111. Lu, S.-L. et al. Randomized multi-reader evaluation of automated detection and segmentation of brain tumors in stereotactic radiosurgery with deep neural networks. *Neuro-Oncology* **23**, 1560–1568 (2021).
112. Willemink, M. J. et al. Preparing medical imaging data for machine learning. *Radiology* **295**, 4–15 (2020).
113. Bauknecht, H.-C. et al. Intra- and interobserver variability of linear and volumetric measurements of brain metastases using contrast-enhanced magnetic resonance imaging. *Investig. Radiol.* **45**, 49–56 (2010).
114. Cho, N. S., Hagiwara, A., Sanvito, F. & Ellingson, B. M. A multi-reader comparison of normal-appearing white matter normalization techniques for perfusion and diffusion MRI in brain tumors. *Neuroradiology* **65**, 559–568 (2023).
115. Covert, E. C. et al. Intra- and inter-operator variability in MRI-based manual segmentation of HCC lesions and its impact on dosimetry. *EJNMMI Phys.* **9**, 90 (2022).
116. Fedorov, A. et al. National Cancer Institute Imaging Data Commons: toward transparency, reproducibility, and scalability in imaging artificial intelligence. *RadioGraphics* **43**, e230180 (2023).
117. McMahan, H. B., Moore, E., Ramage, D., Hampson, S. & Arcas, B. A. Y. Communication-efficient learning of deep networks from decentralized data. In *Proc. 20th International Conference on Artificial Intelligence and Statistics (AISTATS)* (eds Singh, A., and Zhu, J.), Vol. 5 *Proceedings of Machine Learning Research* (2017).
118. Kumar, R. et al. Privacy-preserving blockchain-based federated learning for brain tumor segmentation. *Comput. Biol. Med.* **177**, 108646 (2024).
119. Retraction of: A novel fully automated MRI-based-deep-learning method for classification of IDH mutation status in brain gliomas. *Neuro-Oncology* **25**, 1197–1197 (2023).
120. Gidwani, M. et al. Inconsistent partitioning and unproductive feature associations yield idealized radiomic models. *Radiology* **307**, e220715 (2023).
121. Lacson, R., Eskian, M., Licaros, A., Kapoor, N. & Khorasani, R. Machine learning model drift: predicting diagnostic imaging follow-up as a case example. *J. Am. College Radiol.* **19**, 1162–1169 (2022).
122. Rahmani, K. et al. Assessing the effects of data drift on the performance of machine learning models used in clinical sepsis prediction. *Int. J. Med. Inform.* **173**, 104930 (2023).
123. Ricci Lara, M. A., Echeveste, R. & Ferrante, E. Addressing fairness in artificial intelligence for medical imaging. *Nat. Commun.* **13**, 4581 (2022).
124. Xu, Z. et al. Addressing fairness issues in deep learning-based medical image analysis: a systematic review. *npj Digit. Med.* **7**, 286 (2024).
125. Obermeyer, Z., Powers, B., Vogeli, C. & Mullainathan, S. Dissecting racial bias in an algorithm used to manage the health of populations. *Science* **366**, 447–453 (2019).
126. Yang, Y., Zhang, H., Gichoya, J. W., Katabi, D. & Ghassemi, M. The limits of fair medical imaging AI in real-world generalization. *Nat. Med.* **30**, 2838–2848 (2024).
127. Gichoya, J. W. et al. AI recognition of patient race in medical imaging: a modelling study. *Lancet Digit. Health* **4**, e406–e414 (2022).
128. Geis, J. R. et al. Ethics of artificial intelligence in radiology: summary of the Joint European and North American Multisociety Statement. *Radiology* **293**, 436–440 (2019).
129. Keane, P. A. & Topol, E. J. AI-facilitated health care requires education of clinicians. *The Lancet* **397**, 1254 (2021).
130. Dratsch, T. et al. Automation bias in mammography: the impact of artificial intelligence BI-RADS suggestions on reader performance. *Radiology* **307**, e222176 (2023).
131. Atad, M. et al. Counterfactual explanations for medical image classification and regression using diffusion autoencoder. *Mach. Learn. Biomed. Imaging* **2**, 2103–2125 (2024).
132. Bakas, S. et al. The University of Pennsylvania glioblastoma (UPenn-GBM) cohort: advanced MRI, clinical, genomics, & radiomics. *Sci. Data* **9**, 453 (2022).
133. Scarpace, L. et al. *The Cancer Genome Atlas Glioblastoma Multiforme Collection (TCGA-GBM)* [Dataset]. The Cancer Imaging Archive (2016).
134. Cheng, J. et al. Enhanced performance of brain tumor classification via tumor region augmentation and partition. *PLoS ONE* **10**, e0140381 (2015).
135. Ramakrishnan, D. et al. *A Large Open Access Dataset of Brain Metastasis 3D Segmentations on MRI with Clinical and Imaging Feature Information* [Dataset]. The Cancer Imaging Archive (2023).
136. Pedano, N. et al. *The Cancer Genome Atlas Low Grade Glioma Collection (TCGA-LGG)* [Dataset]. The Cancer Imaging Archive (2016).
137. Wang, L.-B. et al. Proteogenomic and metabolomic characterization of human glioblastoma. *Cancer Cell* **39**, 509–528.e20 (2021).
138. Puchalski, R. B. et al. An anatomic transcriptional atlas of human glioblastoma. *Science* **360**, 660–663 (2018).

Acknowledgements

C.P.B. acknowledges the support of the Rappaport Foundation MGH Research Scholar program. E.R.G. acknowledges the support of NCI R01CA244975.

Author contributions

C.P.B. conceived of and planned this review. F.J.D. and C.P.B. reviewed relevant literature. J.B.P., E.R.G., and J.K.-C. provided critical feedback to help shape the review. F.J.D. took the lead in writing the manuscript. F.J.D., J.B.P., E.R.G., J.K.-C., and C.P.B. contributed to manuscript writing and review. C.P.B. supervised the project.

Competing interests

J.K.-C. acts as a consultant for Siloam Vision and receives institutional research grants from GE HealthCare Technologies, Inc. and Genentech, Inc. Other authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to Christopher P. Bridge.

Reprints and permissions information is available at http://www.nature.com/reprints