Review

Convolutional Neural Network Techniques for Brain Tumor Classification (from 2015 to 2021): Review, Challenges, and Future Perspectives

Yuting Xie [1,†], Fulvio Zaccagna [1,2,†], Leonardo Rundo [3], Claudia Testa [2,4], Raffaele Agati [5], Raffaele Lodi [1,6], David Neil Manners [1,‡,*] and Caterina Tonon [1,2,‡]

[1] Department of Biomedical and Neuromotor Sciences, University of Bologna, 40126 Bologna, Italy; yuting.xie2@unibo.it (Y.X.); fulvio.zaccagna@unibo.it (F.Z.); raffaele.lodi@unibo.it (R.L.); caterina.tonon@unibo.it (C.T.)
[2] IRCCS Istituto delle Scienze Neurologiche di Bologna, Functional and Molecular Neuroimaging Unit, Bellaria Hospital, 40139 Bologna, Italy; claudia.testa@unibo.it
[3] Department of Information and Electrical Engineering and Applied Mathematics, University of Salerno, 84084 Fisciano, SA, Italy; lrundo@unisa.it
[4] Department of Physics and Astronomy, University of Bologna, 40127 Bologna, Italy
[5] IRCCS Istituto delle Scienze Neurologiche di Bologna, Programma Neuroradiologia con Tecniche ad elevata complessità, Bellaria Hospital, 40139 Bologna, Italy; raffaele.agati@isnb.it
[6] IRCCS Istituto delle Scienze Neurologiche di Bologna, Bellaria Hospital, 40139 Bologna, Italy
* Correspondence: davidneil.manners@unibo.it
† These authors contributed equally to this work.
‡ These authors contributed equally to this work.

Abstract: Deep learning has shown remarkable results in every field, especially in the biomedical field, due to its ability to exploit large-scale datasets. A convolutional neural network (CNN) is a widely used deep learning approach to solve medical imaging problems. Over the past few years, many studies have focused on CNN-based techniques for brain tumor diagnosis. There are, however, still some critical challenges that CNNs face towards clinic application. This study presents a comprehensive review of current literature that involves CNN architectures for brain tumor classification. We compare the key achievements in the performance evaluation metrics of the applied classification algorithms. In addition, this review assesses the clinical effectiveness of the included studies to elaborate on the limitations and directions of this area for future work. No review focusing on the clinical effectiveness of previous works in this field has been published. We believe that this study has the potential to elevate the application of CNN-based deep learning methods in clinical practice and also can be a quick reference for biomedical researchers who are interested in this field.

Keywords: deep learning; convolutional neural network; brain tumor classification; clinical application

1. Introduction

Brain tumors are a heterogenous group of common intracranial tumors causing significant mortality and morbidity [1,2]. Malignant brain tumors are among the most aggressive and deadly neoplasms in people of all ages, with mortality rates of 5.4/100,000 men and 3.6/100,000 women per year between 2014 and 2018 [3]. According to the 2021 World Health Organization (WHO) Classification of Tumors of the Central Nervous System, brain tumors are classified into 4 grades (1 to 4) of increasingly aggressive malignancy and worsening prognosis. Indeed, in clinical practice, tumor type and grade influence treatment choice. Within WHO Grade 4 tumors, glioblastoma is the most aggressive primary brain tumor, with median survival after diagnosis of just 12–15 months [4].

Pathological assessment of tissue samples is the reference standard for tumor diagnosis and grading. However, a non-invasive tool capable of accurately classifying tumor type and inferring grade would be highly desirable [5]. Although there are several non-invasive imaging modalities that can visualize brain tumors, i.e., Computed Tomography (CT), Positron Emission Tomography (PET), and Magnetic Resonance Imaging (MRI), the last of them remains the standard of care in clinical practice [6]. MRI conveys information on the lesion location, size, extent, features, relationship with the surrounding structures, and associated mass effect [6]. Beyond structural information, MRI can also assess microstructural features, such as lesion cellularity [7], microvascular architecture [8], and perfusion [9]. Advanced imaging techniques may demonstrate many aspects of tumor heterogeneity related to type, aggressiveness, and grade; however, they are limited in assessing mesoscopic changes that predate macroscopic ones [10]. Many molecular imaging techniques have recently been developed to better reveal and quantify heterogeneity, permitting a more accurate characterization of brain tumors. However, applying this wealth of new information may benefit from more sophisticated and potentially partially automated tools for image analysis [10].

Computer-aided detection and diagnosis (CADe and CADx, respectively), which refer to software that combines artificial intelligence and computer vision to analyze radiological and pathology images, have been developed to help radiologists diagnose human disease in several body districts, such as applications including colorectal polyp detection and segmentation [11,12] and lung cancer classification [13,14].

Machine learning has vigorously accelerated the development of CAD systems [15]. One of the most recent applications of machine learning in CAD is classifying objects of interest, such as lesions, into specific classes based on input features [16-19]. In machine learning, various image analysis tasks can be performed by finding or learning informative features that successfully describe the regularities or patterns in data. However, conventionally, meaningful or task-relevant features are mainly designed by human experts based on their knowledge of the target domain, making it challenging for those without domain expertise to leverage machine learning techniques. Furthermore, traditional machine learning methods can only detect superficial linear relationships, while the biology underpinning living organisms is several orders of magnitude more complex [20].

Deep learning [21], inspired by an understanding of neural networks within the human brain, has achieved unprecedented success in facing the challenges mentioned above by incorporating the feature extraction and selection step into the training process. Generically, deep learning models are represented by a series of layers, each formed by a weighted sum of elements in the previous layer, the first represents the data and the last the output or solution. Multiple layers enable complicated mapping functions to be reproduced, allowing deep learning models to solve very challenging problems while typically needing less human intervention than traditional machine learning. Deep learning currently outperforms alternative machine learning approaches [22] and, for the past few years, has been widely used for a variety of tasks in medical image analysis [23].

A convolutional neural network (CNN) is a deep learning approach that has frequently been applied to medical imaging problems. It overcomes the limitations of previous deep learning approaches because its architecture allows it to automatically learn features important for the problem given a training corpus of sufficient variety and quality [24]. Recently, CNNs have gained popularity for brain tumor classification due to their outstanding performance with very high accuracy in a research context [25-29].

Despite the growing interest in CNN-based CADx within the research community, translation into daily clinical practice has yet to be achieved due to obstacles such as the lack of an adequate amount of reliable data for training algorithms and imbalances within datasets used for multi-class classification [30,31], among others. Several reviews [31-33] have been published in this regard, summarizing the classification methods and key achievements and pointing out some limitations in previous studies, but as yet, none of them have focused on deficiencies regarding clinical adoption or attempted to determine future research directions required to promote the application of deep learning models in clinical practice. For these reasons, the current review considers the key limitations and obstacles regarding the clinical applicability of studies in brain tumor classification by CNN algorithms and how to translate CNN-based CADx technology into better clinical decision-making.

In this review, we explore current evidence on using CNN-based deep learning for brain tumor classification published between 2015 and 2021. The objectives of the review were three-fold: to (1) review and analyze article characteristics and the impact of CNN methods applied to MRI for glioma classification, (2) explore the limitations of current research and the gaps in the bench-to-bedside translation, and (3) find directions for future research in this field. This review was designed to answer the following research questions: How has deep learning been applied to process MR images for glioma classification? What level of impact have papers in this field achieved? How can the translational gap be bridged to deploy deep learning algorithms in clinical practice?

The review is organized as follows. Section 2 introduces the methods used to search and select literature related to the focus of the review. Section 3 presents the general steps of CNN-based deep learning methods for brain tumor classification, and Section 4 introduces relevant primary studies, with an overview of their datasets, preprocessing techniques, and computational methods for brain tumor classification, and presents a quantitative analysis of the covered studies. Furthermore, we introduce the factors that may directly or indirectly degrade the performance and the clinical applicability of CNN-based CADx systems and give an overview of included studies with reference to the degrading factors. Section 5 presents the comparison between studies, and finally, Section 6 summarizes limitations and research trends and suggests directions for further improvements.

2. Materials and Methods

2.1. Article Identification

In this review, we identified preliminary sources using two online databases, PubMed and Scopus. The search queries used for interrogating each database are described in Table 1. The filter option for the publication year (2015–2021) was selected so that only papers in the chosen period are fed into the screening process. Searches were conducted on 04/02/2022. PubMed generated 158 results, and Scopus yielded 265 results.

Table 1. The search queries used for interrogating PubMed and Scopus databases.

PubMed /Scopus	(deep learning OR deep model OR artificial intelligence OR artificial neural network OR autoencoder OR generative adversarial network) OR convolutional OR (neural network) OR neural network OR deep model OR convolutional)	AND
	(brain tumor OR glioma OR brain cancer OR glioblastoma OR astrocytoma OR oligodendroglioma OR ependymoma)	AND
	(classification OR grading OR classify)	AND
	(MRI OR Magnetic Resonance OR MR images OR radiographic OR radiology)	IN
	Title/Abstract	

2.2. Article Selection

Articles were selected for final review using a three-stage screening process based on a series of inclusion and exclusion criteria. After removing duplicate records generated using two databases, articles were first screened based on the title alone. The abstract was then assessed, and finally, the full articles were checked to confirm eligibility. The entire screening process was conducted by one author (Y.T.X). In cases of doubt, records were reviewed by other authors (D.N.M, C.T), and the decision regarding inclusion was arrived at by consensus.

Inclusion criteria were:

- original research articles published in a peer-reviewed journal with full-text access offered by the University of Bologna;
- involved the use of any kind of MR images;
- published in English;
- and concerned with the application of CNN deep learning techniques for brain tumor classification.

Included articles were limited to those published from 2015 to 2021 to focus on deep learning methodologies. A study was defined here as one that employed a CNN-based deep learning algorithm to classify brain tumors and involved the use of one or more of the following performance metrics: accuracy, the area under the receiver operating characteristics curve, sensitivity, specificity, F_1 score.

Exclusion criteria were:

- review article;
- book or book chapter;
- conference paper or abstract;
- short communications or case reports;
- unclear description of data;
- no validation performed.

If a study involved the use of a CNN model for feature extraction but traditional machine learning techniques for the classification task, it was excluded. Studies that used other deep learning networks, for example, artificial neural networks (ANNs), generative adversarial networks (GANs), or autoencoders (AEs), instead of CNN models were excluded. Studies using multiple deep learning techniques, including CNNs, were included in this study, while only the performance of CNNs will be reviewed.

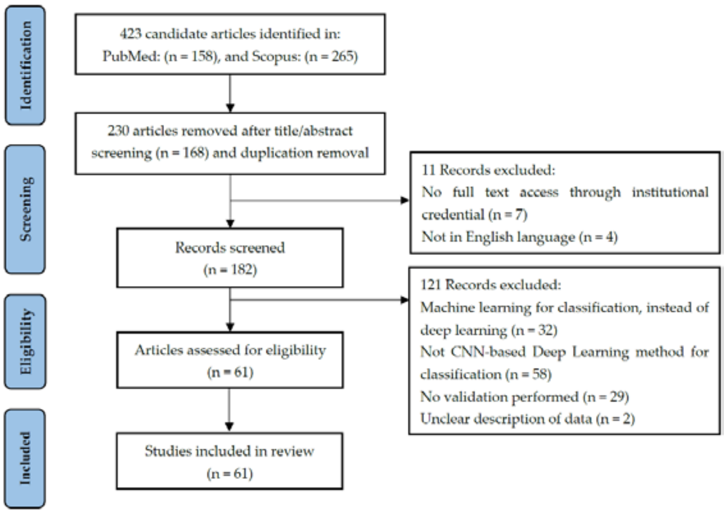

Identification
Screening
Eligibility
Included
423 candidate articles identified in: PubMed: (n = 158), and Scopus: (n = 265)
230 articles removed after title/abstract screening (n = 168) and duplication removal
11 Records excluded:
No full text access through institutional credential (n = 7)
Not in English language (n = 4)
Records screened (n = 182)
121 Records excluded:
Machine learning for classification, instead of deep learning (n = 32)
Not CNN-based Deep Learning method for classification (n = 58)
No validation performed (n = 29)
Unclear description of data (n = 2)
Articles assessed for eligibility (n = 61)
Studies included in review (n = 61)

Figure 1. The PRISMA flowchart of this review.

Figure 1 reports the numbers of articles screened after exclusion at each stage following the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines [34]. A review of 61 selected papers is presented in this paper. All articles cover the aspect of classification of brain tumors using CNN-based deep learning.

3. Literature Review

This section presents a detailed overview of the research papers dealing with brain tumor classification using CNN-based deep learning techniques published during the period from 2015 to 2021. This section is formulated as follows: Section 3.1 presents a brief overview of the general methodology adopted in the majority of the papers for the classification of brain MRI images using CNN algorithms. Section 3.2 presents a description of the popular publicly available datasets that have been used in the research papers reviewed in the form of a Table. Section 3.3 introduces the commonly applied preprocessing methods used in the reviewed studies. Finally, Section 3.4 gives a brief overview of the performance metrics that provide evidence about the credibility of a specific classification algorithm model.

3.1. Basic Architecture of CNN-Based Methods

Recently, deep learning has shown outstanding performance in medical image analysis, especially in brain tumor classification. Deep learning networks have achieved higher accuracy than classical machine learning approaches [22]. In Deep Learning, CNN has achieved significant recognition for its capacity to automatically extract deep features by adapting to small changes in the images [24]. Deep features are those derived from other features that are relevant to the final model output.

The architecture of a typical deep CNN-based brain tumor classification frame is described in Figure 2. To train a CNN-based deep learning model with tens of thousands of parameters, a general rule of thumb is to have at least about 10 times the number of samples as parameters in the network for effective generalization of the problem [35]. Overfitting may occur during the training process if the training dataset is not sufficiently large [36]. Therefore, many studies [37-41] use 2D brain image slices extracted from 3D brain MRI volumes to solve this problem, which increases the number of examples within the initial dataset and mitigates the class imbalance problem. In addition, it has the advantage of reducing the input data dimension and reducing the computational burden of training the network.

Data augmentation is another effective technique for increasing both the amount and the diversity of the training data by adding modified copies of existing data with commonly used morphological techniques, such as rotation, reflection (also referred to as flipping or mirroring), scaling, translation, and cropping [41,42]. Such strategies are based on the assumption that the size and orientation of image patches do not yield robust features for tumor classification.

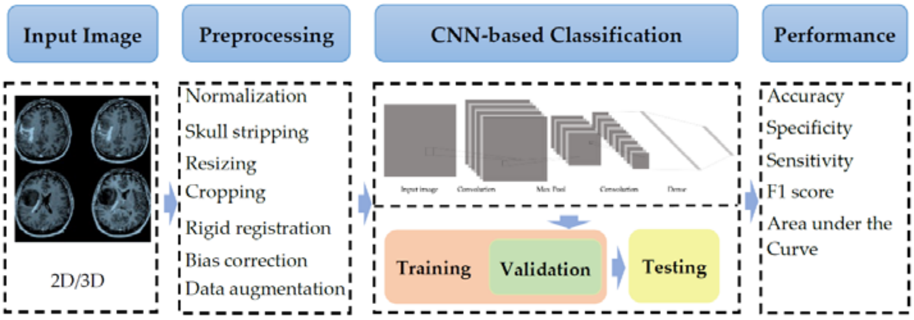


Figure 2. The basic workflow of a typical CNN-based brain tumor classification study with four high-level steps. Step 1. Input Image: 2D or 3D Brain MR samples are fed into the classification model; Step 2. Preprocessing: several preprocessing techniques are used to remove the skull, normalize images, resize images, and augment the number of training examples; Step 3. CNN Classification: The preprocessed dataset is propagated into the CNN model, involving training, validation, and testing process; Step 4. Performance evaluation: Evaluation of the classification performance of a CNN algorithm with accuracy, specificity, F_1 score, area under the curve, and sensitivity metrics.

In deep learning, overfitting is also a common problem when the learning capacity is so large that the network will learn spurious features instead of meaningful patterns [36]. A validation set can be used in the training process to avoid overfitting and obtain a stable performance of the brain tumor classification system on future unseen data in clinical practice. The validation set provides an unbiased evaluation of a classification model on the training data set while tuning the model's hyperparameters during the training process [43]. In addition, validation datasets can be used for regularization by early stopping when the error on the validation data set increases, which is a sign of overfitting to the training data [36,44]. Therefore, in the article selection process, we excluded the articles that omitted validation during the training process.

Evaluation of the classification performance of a CNN algorithm is an essential part of a research study. Accuracy, specificity, F_1 score (also known as the Dice similarity coefficient) [45], the area under the curve, and sensitivity are important metrics to assess the classification model's performance and compare with similar works in the field.

3.2. Datasets

A large training dataset is required to create an accurate and trustworthy deep learning-based classification system for brain tumor classification. In the current instance, this usually comprises a set of MR image volumes, and for each, a classification label is generated by a domain expert such as a neuroradiologist. In the literature reviewed, several datasets have been used for brain tumor classification, targeting both binary tasks [25,37,38,42] and multiclass classification tasks [22,29,46-48]. Table 2 briefly lists some of the publicly accessible databases that have been used by the research work reviewed in this paper, including the MRI sequences included, the size, the classes, the unbiased Gini Coefficient, and the web address of the online repository of the specific dataset.

The Gini coefficient (G) [49] is a property of a distribution that measures its difference from uniformity. It can be applied to categorical data in which classes are sorted by prevalence. Its minimum value is zero if all classes are equally represented, and its maximum varies between 0.5 for a two-class distribution to an asymptote of 1 for many classes. The unbiased Gini coefficient divides G by the maximum value for the number of classes present and takes values in the range 0-1. The maximum value for a distribution with n classes is (n-1)/n. Values of the unbiased Gini coefficient were calculated using R package DescTools [49]. Table 2 allows to appreciate the characteristic of public datasets in terms of balance of samples for the available classes of tumors (unbiased Gini coefficient), although taking under control the total number of samples in the datasets (Column "Size")

Table 2. An overview of publicly available datasets.

Datasets Name	Available Sequences	Size	Classes	Unbiased Gini Coefficient	Source
TCGA-GBM	T_1w, ceT_1w, T_2w, FLAIR	199 patients	N/D	N/D	[50]
TCGA-LGG	T_1w, ceT_1ce, T_2w, FLAIR	299 patients	N/D	N/D	[51]
Brain tumor dataset from Figshare (Cheng et al., 2017)	ceT_1w	233 patients (82 MEN, 89 Glioma, 62 PT), 3064 images (708 MEN, 1426 Glioma, 930 PT)	Patients (82 MEN, 89 Glioma, 62 PT), images (708 MEN, 1426 Glioma, 930 PT)	0.116 (patients), 0.234 (images)	[52]
Kaggle (Navoneel et al., 2019)	No information given	253 images (98 normal, 155 tumorous)	98 normal, 155 tumorous	0.225	[53]
REMBRANDT	T_1w, T_2w, FLAIR, DWI	112 patients (30 AST-II, 17 AST-II, 14 OLI-II, 7 OLI-III, 44 GBM)	30 AST-II, 17 AST-II, 14 OLI-II, 7 OLI-III, 44 GBM	0.402	[54]

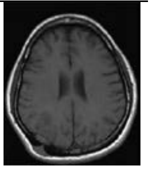
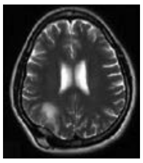
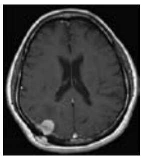
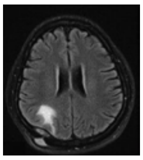
BraTS	T_1w, ceT_1w, T_2w, FLAIR	2019: 335 patients (259 HGG, 76 LGG);
		2018: 284 patients (209 HGG, 75 LGG);
		2017: 285 patients (210 HGG, 75 LGG);
		2015: 274 patients (220 HGG, 54 LGG)
			2019: 259 HGG, 76 LGG;
			2018: 209 HGG, 75 LGG;
			2017: 210 HGG, 75 LGG;
			2015: 220 HGG, 54 LGG
				0.546 (2019);
				0.472 (2018);
				0.474 (2017);
				0.606 (2015)
					[55]

ClinicalTrials.gov (Liu et al., 2017)	T_1w, ceT_1w, T_2w, FLAIR	113 patients (52 LGG, 61 HGG)	52 LGG, 61 HGG	0.080	[56]
CPM-RadPath 2019	T_1w, ceT_1w, T_2w, FLAIR	329 patients	N/D	N/D	[57]
IXI dataset	T_1w, T_2w, DWI	600 normal images	N/D	N/D	[58]
RIDER	T_1w, T_2w, DCE-MRI, ce-FLAIR	19 GBM patients (70220 images)	70,220 images	N/D	[59]
Harvard Medical School Data	T_2w	42 patients (2 normal, 40 tumor), 540 images (27 normal, 513 tumorous)	Patients (2 normal, 40 tumorous), images (27 normal, 513 tumorous)	0.905 (patients), 0.900 (images)	[60]

Among the public datasets, the dataset from Figshare provided by Cheng [52] is the most popular dataset that has been widely used for brain tumor classification. BraTS, referring to the Multimodal Brain Tumor Segmentation Challenge (a well-known challenge that has taken place every year since 2012), is another often-used source of datasets for testing brain tumor classification methods. The provided data are pre-processed, co-registered to the same anatomical template, interpolated to the exact resolution (1 mm^3), and skull stripped [52].

Most MR techniques can generate high-resolution images, while different imaging techniques show distinct contrast, are sensitive to specific tissues or fluid regions and highlight relevant metabolic or biophysical properties of brain tumors [61]. The datasets listed in Table 2 collect one or more MRI sequences, such as T1-weighted (T1w), T2-weighted (T2w), contrast-enhanced T1-weighted (ceT1w), fluid-attenuated inversion recovery (FLAIR), diffusion-weighted imaging (DWI), and dynamic contrast-enhanced magnetic resonance imaging (DCE-MRI). Among these, T1w, T2w, ceT1w, and FLAIR sequences are widely used both in research and clinical practice for brain tumor classification. Each sequence is distinguished by a particular series of radiofrequency pulses and magnetic field gradients, resulting in images with a characteristic appearance [61]. Table 3 lists the imaging configurations and the main clinical distinctions of T$_1$w, T$_2$w, ceT$_1$w, and FLAIR with information retrieved from [61-64].

Table 3. The imaging configurations and main clinical distinctions of T$_1$w, T$_2$w, ceT$_1$w, and FLAIR.

Sequence	Sequence Characteristics	Main clinical distinctions	Example[*]
T$_1$w	Uses short TR and TE [61]	● Lower signal for more water content [63] such as in edema, tumor, inflammation, infection, or chronic hemorrhage [63] ● Higher signal for fat [63] ● Higher signal for subacute hemorrhage [63]	
T$_2$w	Uses long TR and TE [61]	● Higher signal for more water content, as in edema, tumor, infarction, inflammation, infection, subdural collection [63] ● Lower signal for fat [63] ● Lower signal for fibrous tissue [63]	
ceT$_1$w	Uses the same TR and TE as T$_1$w, employs contrast agents [61]	● Higher signal for areas of breakdown in the blood-brain barrier that indicate induced inflammation [62]	
FLAIR	Uses very long TR, TE and the inversion time that nulls the signal from fluid [64]	● Highest signal for abnormalities [62] ● Highest signal for gray matter [64] ● Lower signal for cerebrospinal fluid [64]	

[*]Pictures from [65]. TR, repetition time. TE, echo time.

3.3. Preprocessing

3.3.1. Normalization

The dataset fed into the CNN models may be collected with different clinical protocols and various scanners from multiple institutions. The dataset may consist of MR images of different intensities because the MR image intensities are not consistent across different MR scanners [66]. In addition, the intensity values of MR images are sensitive to the acquisition condition [67]. Therefore, input data should be normalized to minimize the influence of differences between scanners and scanning parameters. Otherwise, any CNN network created will be ill-conditioned.

3.3.2. Skull Stripping

MRI images of the brain normally also contain non-brain regions such as the dura mater, skull, meninges, and scalp. Including these parts in the model typically deteriorates its performance on classification tasks. Therefore, in the studies on brain MRI datasets that retain regions of the skull and vertebral column, skull stripping is widely applied as a preprocessing step in brain tumor classification problems to improve performance [22,68,69].

3.3.3. Resizing

Since deep neural networks require inputs of a fixed size, all images need to be resized before feeding them into the CNN classification models [70]. Images larger than the required size can be resized downwards by either cropping background pixels or scaling down using interpolation [70,71].

3.3.4. Image Registration

Image registration is defined as a process that spatially transforms different images into one coordinate system. In brain tumor classification, it is often necessary to analyze multiple images of a patient to improve the treatment plan, while the images may be acquired from different scanners, at different times, and with different viewpoints [72]. Registration is necessary to be able to integrate the data obtained from these different measurements.

3.3.5. Bias Field Correction

The bias field in medical images is an undesirable artifact caused by factors such as the scan position and instrument used and other unknown issues [73]. This artifact is characterized by differences in brightness across the image and can significantly degrade the performance of many medical image analysis techniques. Therefore, a preprocessing step is needed to correct the bias field signal before submitting corrupted MR images to a CNN classification model.

3.3.6. Data Augmentation

CNN-based classification requires a large amount of data. A general rule of thumb is to have at least about 10 times the number of samples as parameters in the network for effective generalization of the problem [35]. If the database is significantly smaller, overfitting might occur. Data augmentation is one of the foremost preprocessing techniques to subside the imbalance distribution and data scarcity problems. It has been used in many studies that worked on brain tumor classification [22,42,46,47], involving geometrical transformation operations, such as rotation, reflection (also referred to as flipping or mirroring), scaling, translation, and cropping.

Recently, well-established data augmentation techniques are being supplemented by automatic methods using deep learning approaches. For example, the authors in [41] proposed a progressively growing generative adversarial network (PGGAN) augmentation model to help overcome the shortage of images needed for the CNN classification model. However, such methods are rare in the literature reviewed.

3.4. Performance Measures

Evaluation of the classification performance of a CNN algorithm is an essential part of a research study. Here we outline the evaluation metrics most commonly encountered in the brain tumor classification literature, namely accuracy, precision, sensitivity, F1 score, and the area under the curve.

In a classification task, true positive (TP) represents an image that is correctly classified into the positive class according to the ground truth. Similarly, a true negative is an outcome where the model correctly classifies the negative class. On the other hand, false positive (FP) is an outcome where the model incorrectly classifies an image into the positive class while the ground truth is negative. False negative (FN) is an outcome where the model incorrectly classifies the image of a positive class.

3.4.1. Accuracy

Accuracy (ACC) is a metric that measures the performance of a model correctly classifying the classes in a given dataset, given as the percentage of total correct classifications divided by the total number of images.

$$ACC = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

3.4.2. Specificity

Specificity (SPE) represents the proportion of correctly classified negative samples to all negatives identified in the data.

$$SPE = \frac{TN}{TN + FP} \tag{2}$$

3.4.3. Precision

Precision (PRE) represents the ratio of true positives to all identified positives.

$$PRE = \frac{TP}{TP + FP} \tag{3}$$

3.4.4. Sensitivity

Sensitivity (SEN) measures the ability of the classification models to identify positive samples. It represents the ratio of true positives to total (actual) positives in the data.

$$SEN = \frac{TP}{TP + FN} \tag{4}$$

3.4.5. F1 Score

F1 score [45] is one of the most popular metrics considering both precision and recall. It can be used to assess the performance of classification models with class imbalance problems [74], which considers the number of prediction errors that the model makes and looks at the type of errors that are made. It is higher if there is a balance between PRE and SEN.

$$F1\text{ score=}2\frac{PRE * SEN}{PRE + SEN} \tag{5}$$

3.4.6. Area Under the Curve

The area under the curve (AUC) measures the entire two-dimensional area underneath the ROC curve from (0, 0) to (1, 1). It is the measure of the ability of a classifier to distinguish between classes.

4. Results

This section gives an overview of the research papers working on brain tumor classification using CNN techniques. Section 4.1 presents a quantitative analysis of the number of articles published from 2015 to 2021 on deep learning and CNN in brain tumor classification, and the usage of different CNN Algorithms applied in the studies covered. Then, Section 4.2 introduces the factors that may directly or indirectly degrade the performance and the clinical applicability of CNN-based CADx systems. Finally, in Section 4.3, an overview of the included studies will be given with reference to the degrading factors introduced in Section 4.2.

4.1. Quantitative Analysis

As mentioned in the introduction, many CNN models have been used to classify the MR images of brain tumor patients. They overcome the limitations of earlier deep learning approaches and have gained popularity among researchers in brain tumor classification. Figure 3 shows the number of research articles published on PubMed and Scopus in the years from 2015 to 2021 on brain tumor classification by deep learning methods and breaking out CNN-based deep learning techniques; the number of papers related to brain tumor classification using CNN techniques grows rapidly from 2019 and accounts for the majority of the total number in 2020 and 2021. This is because of the high generalizability, stability, and accuracy rate of CNN algorithms.

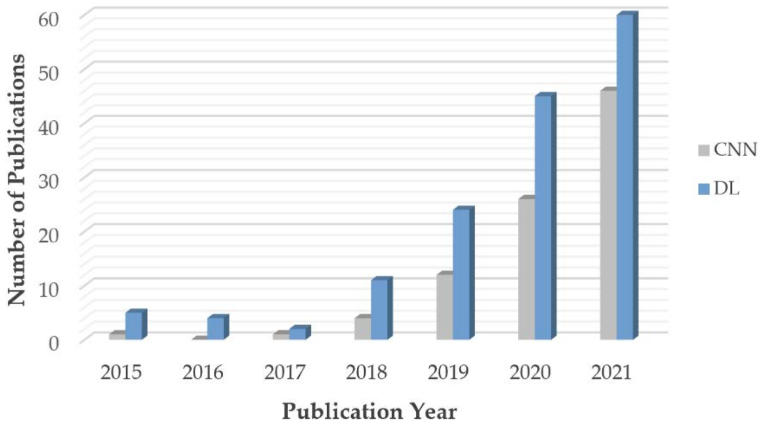


Figure 3. Numbers of articles published from 2015 to 2021.

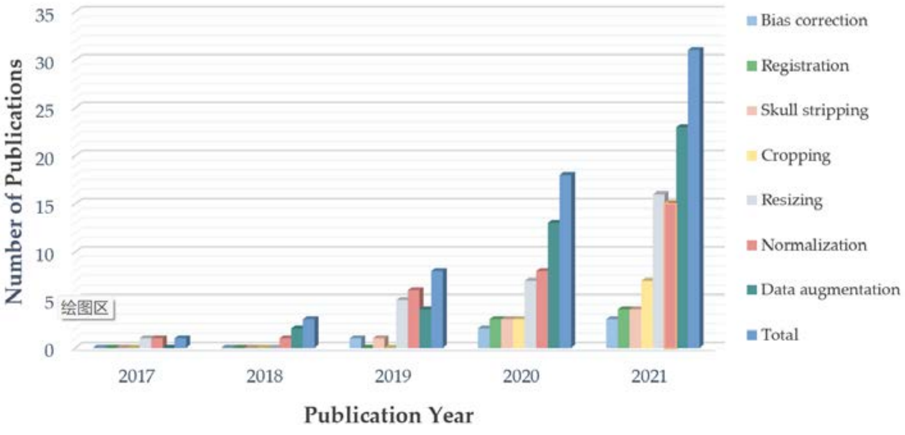


Figure 4. Usage of preprocessing techniques from 2017 to 2021.

Figure 4 shows the usage of most-commonly preprocessing techniques used to address the problems in brain tumor classification, including data augmentation, normalization, resizing, skull stripping, bias field correction, and registration. In this figure, only data from 2017 to 2021 is visualized as there was no article published in 2015 and 2016 that used the preprocessing methods mentioned. Since 2020, data augmentation has been used in the majority of the studies to ease the data scarcity and overfitting problems. However, the bias field problem has not yet been taken seriously, and few studies have included bias field correction in the preprocessing process.

Figure 5 breaks down the usage of publicly available CNN architectures used in the articles included, including custom CNN models, VGG, AlexNet, ResNet, GoogLeNet, DenseNet, and EfficientNet.

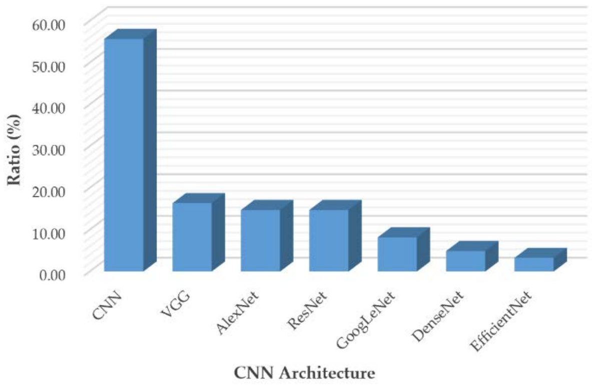


Figure 5. Usage of state-of-the-art CNN models from 2015 and 2021.

AlexNet [75] came out in 2012 and was a revolutionary advancement in deep learning; it improved on traditional CNNs by introducing the composition of consecutively stacked convolutional layers and became one of the best models for image classification. VGG, referring to Visual Geometry Group, was a breakthrough in the world of Convolutional Neural Networks after AlexNet. It is a type of deep CNN architecture with multiple layers proposed by K. Simonyan and A. Zisserman in [76] and was developed to improve model performance by increasing the depth of such CNNs.

GoogLeNet is a deep convolutional neural network with 22 layers based on the Inception architecture; it was developed by researchers at Google [77]. The GoogLeNet addresses most of the problems that large networks face, such as computational expense and overfitting, by employing the Inception module. This module can use max pooling and three varied sizes of filters (1x1, 3x3, 5x5) for convolution in a single image block; such blocks are then concatenated and passed onto the next layer. An extra 1x1 convolution before the 3x3 and 5x5 layers can be added to the neural network to make the process even less computationally expensive [77]. ResNet stands for Deep Residual Network. It is an innovative convolutional neural network proposed in [78]. ResNet makes use of residual blocks to improve the accuracy of the models. A Residual block is a skip-connection block typically with double- or triple-layer skips that contain nonlinearities (ReLU) and batch normalization in between; it can help reduce the problem of vanishing gradients or to mitigate the accuracy saturation problem [78]. DenseNet, which stands for Dense Convolutional Network, is a type of convolutional neural network that utilizes dense connections between layers. DenseNet was developed mainly to improve the declined accuracy caused by the vanishing gradient in neural networks [79]. Additionally, those CNNs take in images of 224×224 pixels. Therefore, for brain tumor classification, the authors need to center crop a 224x224 patch in each image to keep the input image size consistent.

Convolutional Neural Networks are commonly built at a fixed resource budget. When more resources are available, the depth, width, and resolution of the model need to be scaled up for better accuracy and efficiency [80]. Unlike previous CNNs, EfficientNet is a novel baseline network that uses a different model scaling technique based on a compound coefficient and neural architecture search methods that can carefully balance network depth, width, and resolution [80].

4.2. Clinical Applicability Degrading Factors

This section introduces the factors that hinder the adoption and development of CNN-based brain tumor classification CADx systems into clinic practice, including data quality, data scarcity, data mismatch, data imbalance, classification performance, research value towards clinic needs, and the Black-Box characteristics of CNN models.

4.2.1. Data Quality

During the MR image acquisition process, both the scanner and external sources may produce electrical noise in the receiver coil, generating image artifacts in the brain MR volumes [66]. In addition, the MR image reconstruction process is sensitive to acquisition conditions, and further artifacts are introduced if the subject under examination moves during the acquisition of a single image [66]. These errors are inevitable and reduce the quality of MR images used to train the networks. As a result, the quality of the training data highly impacts the performance of the CNN models, thus degrading their applicability in real-world clinic adoption.

4.2.2. Data Scarcity

Big data is also one of the biggest challenges that CNN-based CADx systems face today. A large amount of high-quality annotated data is required for building high-performance CNN classification models, while it is a challenge to label a large number of medical images due to the complexity of medical data. When a CNN classification system lacks a quantity of data, overfitting can occur, affecting the generalization capability of the network to handle new data [81].

4.2.3. Data Mismatch

Data mismatch refers to a situation in which a model well-trained in a lab environment may fail to generalize to real-world clinical data. It might be caused by the overfitting of the training set or due to the mismatch between the research images and clinic ones [74]. Studies are at high risk of generalization failure if they omit a validation step.

4.2.4. Class Imbalance

In brain MRI datasets, such as the BraTS 2019 dataset [82], which consists of 210 HGG and 75 LGG patients (unbiased Gini coefficient 0.546, as shown in Table 2), HGG is represented by a much higher percentage of samples than LGG, leading to the so-called class imbalance problems, where inputting all the data into the CNN classifier to build up the learning model will usually lead to a learning bias to the majority class [83].

4.2.5. Research Value towards Clinical Needs

Different brain tumor classification tasks have been studied with CNN-based deep learning techniques during the period from 2015 to 2021, including clinically relevant 2-class classification (normal vs. tumorous[28,38,84,85], HGG vs. LGG [25,37,42,69], LGG-II vs. LGG-III [86], etc.), 3-class classification (normal vs. LGG vs. HGG [22], meningioma (MEN) vs. pituitary tumor (PT) vs. glioma [36,39,46,47], glioblastoma multiforme (GBM) vs. astrocytoma (AST) vs. oligodendroglioma(OLI) [29], etc.), 4-class classification (LGG vs. OLI vs. anaplastic glioma (AG) vs. GBM [68], normal vs. AST-II vs. OLI-III vs. GBM-IV) [22], normal vs. MEN vs. PT vs. glioma [87], etc.), 5-class classification (AST-II vs. AST-III vs. OLI-II vs. OLI-III vs. GBM-IV [22]), and 6-class classification (normal vs. AST-II vs. AST-III vs. OLI-II vs. OLI-III vs. GBM-IV [22]).

Different classification tasks have different difficulty levels both in the research community and clinical practice. The authors in [22] used AlexNet for multi-class classification tasks, including 2-class classification: normal vs. tumor, 3-class classification: normal vs. LGG vs. HGG; 4-class classification: normal vs. AST vs. OLI vs. GBM; 5-class classification: AST-II vs. AST-III vs. OLI-II vs. OLI-III vs. GBM-IV and 6-class classification: normal vs. AST-II vs. AST-III vs. OLI-II vs. OLI-III vs. GBM-IV. The results reported 100% accuracy for the normal vs. tumorous classification. The accuracy for 5-class classification (AST-II vs. AST-III vs. OLI-II vs. OLI-III vs. GBM-IV) was only 87.14%. Similarly, in a recent publication [88], the authors utilized the same CNN model for multi-class brain tumor classification. The overall accuracy obtained for normal vs. tumorous classification reached 100%, compared with the lower accuracy of 90.35% obtained for the 4-class classification task (Grade I vs. Grade II vs. Grade III vs. Grade IV) and 86.08% for 5-class classification between AST-II vs. AST-III vs. OLI-II vs. OLI-III vs. GBM.

The goal of research in the field of CADx is to help address existing unmet clinical needs and provide assistance methods and tools for the difficult tasks that human professionals cannot easily handle in clinic practice. It is observed that CNN-based models have achieved quite a high accuracy for normal/tumorous image classification, while more research is needed to improve the classification performance of more-difficult tasks, especially between 5-class classification AST-II vs. AST-III vs. OLI-II vs. OLI-III vs. GBM and 4-class classification between Grade I vs. Grade II vs. Grade III vs. Grade IV. Therefore, research that uses normal vs. tumorous as their target problem has little clinical value.

4.2.6. Classification Performance

Classification performance, indicating the reliability and trustworthiness of the CADx systems, is one of the most important factors to be considered when translating research findings into clinical practice. It has been shown that CNN techniques performed well in most of the brain tumor classification tasks, such as 2-class classification between normal and tumorous[84,85], HGG and LGG [42][69]) and 3-class classification between normal vs. LGG vs. HGG [22], MEN vs. PT vs. glioma [46,47]. However, the classification performance obtained for more difficult classification tasks, such as 5-class classification between AST-II, AST-III, OLI-II, OLI-III, and GBM remains poor [22][88] and justifies further research.

4.2.7. Black-Box Characteristics of CNN Models

CNN-based deep learning techniques have shown remarkable performance on brain tumor classification. Still, their clinical application is also limited by another factor, the 'Black-Box' problem: even the designers of a CNN model cannot usually explain its internal workings or why it arrived at a specific decision. The lack of explainability impedes the adoption and development of deep learning tools into clinical practice [89].

4.3. Overview of Included Studies

Many research papers have emerged following the wave of enthusiasm for CNN-based deep learning techniques from 2015 to the present time. In this review, 61 research papers are assessed to summarize the effectiveness of CNN algorithms in brain tumor classification and to suggest directions for future research in this field.

Among the included articles, 15 articles use normal/tumorous as their classification target. However, as mentioned in Section 4.2.5, the differentiation between normal and tumorous images is not a difficult task. It has been well solved both in research and clinic practice, thus having little value for clinical application. Therefore, studies that use normal vs. tumorous as their target problem will not be reviewed in the following assessment steps.

Table 4.1 gives an overview of included studies that focus on CNN-based deep learning methods for brain tumor classification, except studies working on normal vs. tumorous classification. Datasets, MRI sequences, size of the dataset, and preprocessing methods are summarized. Table 4.2 summarizes the classification tasks, classification architecture, validation methods, and performance metrics of the reviewed articles.

As introduced in Section 4.2, the major challenge confronting brain tumor classification by CNN techniques in MR images lies in the training data, including challenges caused by the data quality, data scarcity, data mismatch, and data imbalance that hinder the adoption and development of CNN-based brain tumor classification CADx systems into clinic practice. Here we assess several newly published literature to provide a convenient collection of the state-of-the-art techniques used to address these issues and the problems that have not been solved in the studies.

Data augmentation is recognized as the current best solution to the problem caused by the scarcity of data and has been widely utilized in brain tumor classification studies.

Authors in [90] used different data augmentation methods, including rotation, flipping, Gaussian Blur, sharpening, edges detection, embossing, skewing, and shearing to increase the size of the dataset. The proposed system aims to classify between Grade I, Grade II, Grade III, and Grade IV, and the original data consists of 121 images (36 Grade I images, 32 Grade II images, 25 Grade III images, and 28 Grade IV images), and by using data augmentation techniques, 30 new images are generated from each MR image. The proposed model is experimentally evaluated on both augmented and original data. The results show that the overall accuracy after data augmentation reaches 90.67%, greater than the accuracy of 87.38% obtained without augmentation.

In a recent publication by Allah et al. [41], a novel data augmentation method called progressive growing generative adversarial network (PGGAN) was proposed and combined with rotation and flipping methods involving an incremental increase of the size of the model during the training to produce MR images of brain tumors and to help overcome the shortage of images for deep learning training. The brain tumor images were classified using a VGG19 features extractor coupled with a CNN classifier. The accuracy of the combined VGG19 + CNN and PGGAN data augmentation framework achieved an accuracy of 98.54%.

Another approach that helps overcome the problem of data scarcity and can also reduce computational cost and training time is transfer learning. Transfer learning is a hot research topic in machine learning; previously learned knowledge can be transferred to the performance of a new task by fine-tuning a previously generated model with a smaller data set that is more specific to the aim of the study. Transfer learning is usually expressed by using pre-trained models, such as VGG, GoogLeNet, and AlexNet, that have been trained on the large benchmark dataset ImageNet [91].

Table 4.1. Overview of included studies that focus on CNN-based deep learning methods for brain tumor classification, excepting studies work on normal vs. tumorous classification. Datasets, MRI sequences, size of the dataset, and preprocessing methods are summarized.

			Size of Dataset		Pre-processing						Data augmentation						
Author & Year	**Datasets**	**MRI Sequences**	**Patients**	**Images**	**Cropping**	**Normalization**	**Resizing**	**Skull stripping**	**Registration1**	**Other**	**Translation[1]**	**Rotation**	**Scaling[2]**	**Reflection[3]**	**Shearing**	**Cropping**	**Other (X= unspecified)**
Özcan et al. [25] 2021	Private dataset	T_2w/ FLAIR	104 (50 LGG, 54 HGG)	518	x	x				Conversion to BMP		x	x	x	x		
Hao et al. [92] 2021	BraTS 2019	T_1w, ceT_1w, T_2w	335 (259 HGG, 76 LGG)	6700			x	x	x								
Tripathi et al. [93] 2021	1. TCGA-GBM, 2. LGG-1p19qDeletion	T_2w	322 (163 HGG, 159 LGG)	7392 (5088 LGG, 2304 HGG)				x			x	x	x	x		x	
Ge et al. [37] 2020	BraTS 2017	T_1w, ceT_1w, T_2w, FLAIR	285 (210 HGG, 75 LGG)								x			x			
Mzoughi et al. [27] 2020	BraTS 2018	ceT_1w	284 (209 HGG, 75 LGG)			x	x			Contrast enhancement				x			
Yang et al. [42] 2018	ClinicalTrials.gov (NCT026226201)	ceT_1w	113 (52 LGG, 61 HGG)							Conversion to BMP		x	x	x			Histogram equalization, adding noise
Zhuge et al. [94] 2020	1.TCIA-LGG, 2. BraTS 2018	T_1w, T_2w, FLAIR, ceT_1w	315 (210 HGG, 105 LGG)			x			x	Clipping, bias field correction		x	x	x			
Decuyper et al. [69] 2021	1. TCGA-LGG, 2. TCGA-GBM, 3. TCGA-1p19qDeletion, 4. BraTS 2019. 5. GUH dataset	T_1w, ceT_1w, T_2w, FLAIR	738 (164 from TCGA-GBM, 121 from TCGA-LGG, 141 from 1p19qDeletion, 202 from BraTS 2019, 110 from GUH dataset) (398 GBM vs. 340 LGG)			x		x	x	Interpolation		x		x			Elastic transform
He et al. [95] 2021	1.Dataset from TCIA	FLAIR, ceT_1w	214 (106 HGG, 108 LGG)			x	x		x								x
	2. BraTS 2017	FLAIR, ceT_1w	285 (210 HGG, 75 LGG)			x	x		x								x
Hamdaoui et al. [96] 2021	BraTS 2019	T_1w, ceT_1w, T_2w, FLAIR	285 (210 HGG, 75 LGG)	53064 (26532 HGG, 26532 LGG)	x								x	x			
Chikhalikar et al. [97] 2021	BraTS 2015	T_2w, FLAIR	274 (220 HGG, 54 LGG)	521						Contrast enhancement							
Ahmad [98] 2019	BraTS 2015	No info shared		124 (99 HGG, 25 LGG)		x											
Naser et al. [86] 2020	TCGA-LGG	T_1W, FLAIR, ceT_1w	108 (50 Grade II, 58 Grade III)		x	x	x			Padding	x	x	x	x	x		
Allah et al. [41] 2021	Figshare (Cheng et al., 2017)	ceT_1w	233 (as shown in Table 2)	3064 (as shown in Table 2)		x						x		x			PGGAN
Swati et al. [47] 2019	Figshare (Cheng et al., 2017)	ceT_1w	233 (as shown in Table 2)	3064 (as shown in Table 2)		x	x										
Guan et al. [40] 2021	Figshare (Cheng et al., 2017)	ceT_1w	233 (as shown in Table 2)	3064 (as shown in Table 2)		x	x			Contrast enhancement		x		x			

Deepak et al. [36] 2019	Figshare (Cheng et al., 2017)	ceT$_1$w	233 (as shown in Table 2)	3064 (as shown in Table 2)		x	x										
Díaz-Pernas et al. [39] 2021	Figshare (Cheng et al., 2017)	ceT$_1$w	233 (as shown in Table 2)	3064 (as shown in Table 2)		x											Elastic transform
Ismael et al. [46] 2020	Figshare (Cheng et al., 2017)	ceT$_1$w	233 (as shown in Table 2)	3064 (as shown in Table 2)	x		x				x	x	x	x	x		Whitening, brightness manipulation

Table 4.1. (*Continued*)

			Size of Dataset		Pre-processing						Data augmentation						
Author & Year	Datasets	MRI Sequences	Patients	Images	Cropping	Normalization	Resizing	Skull stripping	Registration1	Other	Translation^1	Rotation	Scaling^2	Reflection^3	Shearing	Cropping	Other (X= unspecified)
Alhassan et al. [99] 2021	Figshare (Cheng et al., 2017)	ceT$_1$w	233 (as shown in Table 2)	3064 (as shown in Table 2)		x											
Bulla et al. [100] 2020	Figshare (Cheng et al., 2017)	ceT$_1$w	233 (as shown in Table 2)	3064 (as shown in Table 2)		x	x										
Ghassemi et al. [101] 2020	Figshare (Cheng et al., 2017)	ceT$_1$w	233 (as shown in Table 2)	3064 (as shown in Table 2)		x						x		x			
Kakarla et al. [102] 2021	Figshare (Cheng et al., 2017)	ceT$_1$w	233 (as shown in Table 2)	3064 (as shown in Table 2)		x	x			Contrast enhancement							
Noreen et al. [103] 2021	Figshare (Cheng et al., 2017)	ceT$_1$w	233 (as shown in Table 2)	3064 (as shown in Table 2)		x											
Noreen et al. [104] 2020	Figshare (Cheng et al., 2017)	ceT$_1$w	233 (as shown in Table 2)	3064 (as shown in Table 2)		x											
Kumar et al. [105] 2021	Figshare (Cheng et al., 2017)	ceT$_1$w	233 (as shown in Table 2)	3064 (as shown in Table 2)								x					
Badža et al. [106] 2020	Figshare (Cheng et al., 2017)	ceT$_1$w	233 (as shown in Table 2)	3064 (as shown in Table 2)		x	x					x		x			
Alaraimi et al. [107] 2021	Figshare (Cheng et al., 2017)	ceT$_1$w	233 (as shown in Table 2)	3064 (as shown in Table 2)		x	x				x	x	x	x		x	
Lo et al. [108] 2019	Dataset from TCIA**	ceT$_1$w	130 (30 Grade II, 43 Grade III, 57 Grade IV)			x	x			Contrast enhancement	x	x	x	x		x	
Kurc et al. [109] 2020	Data from TCGA	ceT$_1$w, T$_2$-FLAIR	32 (16 OLI, 16 AST)					x	x	Bias field correction		x				x	
Pei et al. [110] 2020	1. CPM-RadPath 2019, 2. BraTS 2019	T$_1$w, ceT$_1$w, T$_2$w, FLAIR	398 (329 from CPM-RadPath 2019, 69 from BraTS 2019)			x		x	x	Noise reduction		x	x			x	

Ahammed et al. [68] 2019	Private dataset	T_2w	20	557 (130 Grade I, 169 Grade II, Grade III 103, Grade IV 155)				x		Filtering, enhancement	x	x	x	x			
Mohammed et al. [48] 2020	Radiopaedia	No info shared	60 (15 of each class)	1258 (311 EP, 286 normal, 380 MEN, 281 MB)			x			Denoising	x	x	x	x		x	
McAvoy et al. [111] 2021	Private dataset	ceT_1w	320 (160 GBM, 160 PCNSL)	3887 (2332 GBM, 1555 PCNSL)		x	x			Random changes to color, noise sampling				x			
Gilanie et al. [112] 2021	Private dataset	T_1w, T_2w, FLAIR	180 (50 AST-I, 40 AST-II, 40 AST-III, 50 AST-IV)	30240 (8400 AST-I, 6720 AST-II, 6720 AST-III, 8400 AST-IV)		x				Bias field correction		x					
Kulkarni et al. [113] 2021	Private dataset	T_1w, T_2w, FLAIR		200 (100 benign, 100 malignant)						Denoising, contrast enhancement	x	x	x	x	x		

Table 4.1. (*Continued*).

Author & Year	Datasets	MRI Sequences	Size of Dataset		Pre-processing						Data augmentation						
			Patients	Images	Cropping	Normalization	Resizing	Skull stripping	Registration1	Other	Translation[1]	Rotation	Scaling[2]	Reflection[3]	Shearing	Cropping	Other (X= unspecified)
Artzi et al. [114] 2021	Private dataset	T_1w, FLAIR, DTI	158 (22 Normal, 63 PA, 57 MB, 16 EP)	731 (110 Normal, 280 PA, 266 MB, 75 EP)	x		x		x	Background removal, bias field correction		x	x	x			Brightness changes
Gu et al. [29] 2021	1. Figshare (Cheng et al., 2017)	ceT_1w	233 (as shown in Table 2)	3064 (as shown in Table 2)			x										
	2. REMBRANDT	No info shared	130	110020			x										
Rajini [115] 2019	1. IXI dataset, REMBRANDT, TCGA-GBM, TCGA-LGG	No info shared	600 normal images from IXI dataset, 130 patients from REMBRANDT, 200 patients from TCGA-GBM, 299 patients from TCGA-LGG														
	2. Figshare (Cheng et al., 2017)	ceT_1w	233 (as shown in Table 2)	3064 (as shown in Table 2)													
Anaraki et al. [116] 2019	1: IXI dataset, REMBRANDT, TCGA-GBM, TCGA-LGG, private dataset	no info of IXI, ceT_1w from REMBRANDT, TCGA-GBM, TCGA-LGG	600 normal images from IXI dataset, 130 patients from REMBRANDT, 199 patients from TCGA-GBM, 299 patients from TCGA-LGG, 60 patients from private dataset			x	x				x	x	x	x			
	2. Figshare (Cheng et al., 2017)	ceT_1w	233 (as shown in Table 2)	3064 (as shown in Table 2)		x	x				x	x	x	x			

Sajjad et al. [90] 2019	1. Radiopaedia	No info shared		121 (36 Grade I, 32 Grade II, 25 Grade III, 28 Grade IV)		x	x			Denoising, bias field correction		x		x	x		Gaussian blurring, sharpening, embossing, skewing
	2. Figshare (Cheng et al., 2017)	ceT$_1$w	233 (as shown in Table 2)	3064 (as shown in Table 2)		x	x			Denoising, bias field correction		x		x	x		Gaussian blurring, sharpening, embossing, skewing
Wahlang et al. [117] 2020	1. Radiopaedia	FLAIR	11 (2 Metastasis, 6 Glioma, 3 MEN)													x	
	2. BraTS 2017	No info shared	20	3100						Median filtering							
Xiao et al. [87] 2021	1. Private dataset	No info shared		1109 (495 MT, 614 Normal)			x										
	2. Figshare (Cheng et al., 2017)	ceT$_1$w	233 (as shown in Table 2)	3064 (as shown in Table 2)			x										
	3. Brain Tumor Classification (MRI) Dataset from Kaggle	No info shared		3264 (937 MEN, 926 Glioma, 901 PT, 500 Normal)			x										

Table 4.1. (*Continued*).

			Size of Dataset		Pre-processing						Data augmentation						
Author & Year	**Datasets**	**MRI Sequences**	**Patients**	**Images**	**Cropping**	**Normalization**	**Resizing**	**Skull stripping**	**Registration1**	**Other**	**Translation^1**	**Rotation**	**Scaling^2**	**Reflection^3**	**Shearing**	**Cropping**	**Other (X= unspecified)**
Tandel et al. [118] 2021	REMBRANDT	T$_2$w	See 1-4 below	See 1-4 below			x			Converted to RGB		x	x				
			130	1. 2156 (1041 normal, 1091 tumorous)													
			47	2. 557 (356 AST-II, 201 AST-III)													
			21	3. 219 (128 OLI-II, 91 OLI-III)													
			112	4. 1115 (484 LGG, 631 HGG)													
Tandel et al. [22] 2020	REMBRANDT	T$_2$w	112 (30 AST-II, 17 AST-II, 14 OLI-II, 7 OLI-III, 44 GBM)	See 1-5 below				x				x	x				
				1. 2132 (1041 normal, 1091 tumorous)													
				2. 2156 (1041 normal, 484 LGG, 631 HGG)													
				3. 2156 (1041 normal, 557 AST, 219 OLI, 339 GBM)													
				4. 1115 (356 AST-II, 201 AST-III, 128 OLI-II, 91 OLI-III, 339 GBM)													
				5. 2156 (1041 normal, 356 AST-II, 201 AST-III, 128 OLI-II, 91 OLI-III, 339 GBM)													
Ayadi et al. [88] 2021	1. Radiopaedia	No info shared		121 (36 Grade I, 32 Grade II, 25 Grade III, 28 Grade IV)								x		x			Gaussian blurring, sharpening
	2. Figshare (Cheng et al., 2017)	ceT$_1$w	233 (as shown in Table 2)	3064 (as shown in Table 2)													
	3. REMBRANDT	FLAIR, T$_1$w, T$_2$w	130 (47 AST, 21 OLI, 44 GBM, 18 unknown)	See 1-5 below								x		x			Gaussian blurring, sharpening"
				1. 2132 (1041 normal, 1091 tumorous)													
				2. 2156 (1041 normal, 484 LGG, 631 HGG)													

	3. 2156 (1041 normal, 557 AST, 219 OLI, 339 GBM) 4. 1115 (356 AST-II, 201 AST-III, 128 OLI-II, 91 OLI-III, 339 GBM) 5. 2156 (1041 normal, 356 AST-II, 201 AST-III, 128 OLI-II, 91 OLI-III, 339 GBM)

Notes: 1. Rigid registration unless otherwise notes; 2. Translation also referred to as shifting; 3. Scaling also referred to as zooming; 4. Reflection also referred to as flipping or mirroring.

Table 4.2. Overview of included studies that focus on CNN-based deep learning methods for brain tumor classification, excepting studies work on normal vs. tumorous classification. Classification tasks, classification architecture, validation methods, and performance metrics are summarized.

Author & Year	Classification Tasks	Model Architecture	Validation	Performance	ACC%[1]
2 classes					
Özcan et al. [25] 2021	LGG (grade II) vs. HGG (grade IV)	Custom CNN model	5-fold CV	SEN = 98.0%, SPE = 96.3%, F1 score = 97.0%, AUC = 0.989	97.1
Hao et al. [92] 2021	LGG vs. HGG	Transfer learning with AlexNet	No info shared	AUC = 82.89%	
Tripathi et al. [93] 2021	LGG vs. HGG	Transfer learning with Resnet18	No info shared		95.87
Ge et al. [37] 2020	LGG vs. HGG	Custom CNN model	No info shared	SEN = 84.35%, SPE = 93.65%	90.7
Mzoughi et al. [27] 2020	LGG vs. HGG	Multi-scale 3D CNN	No info shared		96.49
Yang et al. [42] 2018	LGG vs. HGG	Transfer learning with AlexNet, GoogLeNet	5-fold CV	AUC = 0.939	86.7
Zhuge et al. [94] 2020	LGG vs. HGG	Transfer learning with ResNet50	5-fold CV	SEN = 93.5%, SPE = 97.2%	96.3
		3D CNN	5-fold CV	SEN = 94.7%, SPE = 96.8%	97.1
Decuyper et al. [69] 2021	LGG vs. GBM	3D CNN	No info shared	SEN = 90.16%, SPE = 89.80%, AUC = 0.9398	90
He et al. [95] 2021	LGG vs. HGG	Custom CNN model	5-fold CV	TCIA: SEN = 97.14%, SPE = 90.48%, AUC = 0.9349	92.86
				BraTS 2017: SEN = 95.24%, SPE = 92%, AUC = 0.952	94.39
Hamdaoui et al. [96] 2021	LGG vs. HGG	Transfer learning with stacking VGG16, VGG19, MobileNet, InceptionV3, Xception, Inception ResNetV2, DenseNet121	10-fold CV	PRE = 98.67%, F1 score = 98.62%, SEN = 98.33%	98.06
Chikhalikar et al. [97] 2021	LGG vs. HGG	Custom CNN model	No info shared		99.46
Ahmad [98] 2019	LGG vs. HGG	Custom CNN model	No info shared		88
Naser et al. [86] 2020	LGG (Grade II) vs. LGG (Grade III)	Transfer learning with VGG16	5-fold CV	SEN = 97%, SPE = 98%	95
Kurc et al. [109] 2020	OLI vs. AST	3D CNN	5-fold CV		80
McAvoy et al. [111] 2021	GBM vs. PCNSL	Transfer learning with EfficientNetB4	No info shared	GBM: AUC = 0.94, PCNSL: AUC = 0.95	

Kulkarni et al. [113] 2021	Benign vs. Malignant	Transfer learning with AlexNet	5-fold CV	PRE = 93.7%, RE = 100%, F1 score = 96.77%	
		Transfer learning with VGG16	5-fold CV	PRE = 55%, RE = 50%, F1 score = 52.38%	
		Transfer learning with ResNet18	5-fold CV	PRE = 78.94%, RE = 83.33%, F1 score = 81.07%	
		Transfer learning with ResNet50	5-fold CV	PRE = 95%, RE = 55.88%, F1 score = 70.36%	
		Transfer learning with GoogLeNet	5-fold CV	PRE = 75%, RE = 100%, F1 score = 85.71%	

Table 4.2. (*Continued*).

Author & Year	Classification Tasks	Model Architecture	Validation	Performance	ACC%[1]
Wahlang et al. [117] 2020	HGG vs. LGG	AlexNet	No info shared		62
		U-Net	No info shared		60
Xiao et al. [87] 2021	MT vs. Normal	Transfer learning with ResNet50	3-fold, 5-fold, 10-fold CV	AUC = 0.9530	98.2
Tandel et al. [118] 2021	1. Normal vs. Tumorous	DL-MajVot (AlexNet, VGG16, ResNet18, GoogleNet, ResNet50)	5-fold CV	SEN = 96.76%, SPE = 96.43%, AUC = 0.966	96.51
	2. AST-II vs. AST-III	DL-MajVot (AlexNet, VGG16, ResNet18, GoogleNet, ResNet50)	5-fold CV	SEN = 94.63%, SPE = 99.44%, AUC = 0.9704	97.7
	3. OLI-II vs. OLI-III	DL-MajVot (AlexNet, VGG16, ResNet18, GoogleNet, ResNet50)	5-fold CV	SEN = 100%, SPE = 100%, AUC = 1	100
	4. LGG vs. HGG	DL-MajVot (AlexNet, VGG16, ResNet18, GoogleNet, ResNet50)	5-fold CV	SEN = 98.33%, SPE = 98.57%, AUC = 0.9845	98.43
Tandel et al. [22] 2020	Normal vs. Tumorous	Transfer learning with AlexNet	Multiple CV (K2, K5, K10)	RE = 100%, PRE = 100%, F1 score = 100%	100
Ayadi et al. [88] 2021	Normal vs. Tumorous	Custom CNN model	5-fold CV		100
3 classes					
Allah et al. [41] 2021	MEN vs. Glioma vs. PT	PGGAN-augmentation VGG19	No info shared		98.54
Swati et al. [47] 2019	MEN vs. Glioma vs. PT	Transfer learning with VGG19	5-fold CV	SEN = 94.25%, SPE = 94.69%, PRE = 89.52%, F1 score = 91.73%	94.82
Guan et al. [40] 2021	MEN vs. Glioma vs. PT	EfficientNet	5-fold CV		98.04
Deepak et al. [36] 2019	MEN vs. Glioma vs. PT	Transfer learning with GoogleNet	5-fold CV		98
Díaz-Pernas et al. [39] 2021	MEN vs. Glioma vs. PT	Multiscale CNN	5-fold CV		97.3
Ismael et al. [46] 2020	MEN vs. Glioma vs. PT	Residual networks	5-fold CV	PRE = 99.0%, RE = 99.0%, F1 score = 99.0%	99
Alhassan et al. [99] 2021	MEN vs. Glioma vs. PT	Custom CNN model	k-fold CV	PRE = 99.6%, RE = 98.6%, F1 score = 99.0%	98.6

Bulla et al. [100] 2020	MEN vs. Glioma vs. PT	Transfer learning with InceptionV3 CNN model	holdout validation, 10-fold CV, stratified 10-fold CV, group 10-fold CV	Under group 10-fold CV: PRE = 97.57%, RE = 99.47%, F1 score = 98.40%, AUC = 0.995	99.82
Ghassemi et al. [101] 2020	MEN vs. Glioma vs. PT	CNN-GAN	5-fold CV	PRE = 95.29%, SEN = 94.91%, SPE = 97.69%, F1 score = 95.10%	95.6
Kakarla et al. [102] 2021	MEN vs. Glioma vs. PT	Custom CNN model	5-fold CV	PRE = 97.41%, RE = 97.42%	97.42
Noreen et al. [103] 2021	MEN vs. Glioma vs. PT	Transfer learning with Inception-v3	K-fold CV		93.31
		Transfer learning with Inception model	K-fold CV		91.63
Noreen et al. [104] 2020	MEN vs. Glioma vs. PT	Transfer learning with Inception-v3	No info shared		99.34
		Transfer learning with DensNet201	No info shared		99.51
Kumar et al. [105] 2021	MEN vs. Glioma vs. PT	Transfer learning with ResNet50	5-fold CV	PRE = 97.20%, RE = 97.20%, F1 score = 97.20%	

Table 4.2. (*Continued*).

Author & Year	Classification Tasks	Model Architecture	Validation	Performance	ACC%[1]
Badža et al. [106] 2020	MEN vs. Glioma vs. PT	Custom CNN model	10-fold CV	PRE = 95.79%, RE = 96.51%, F1 score = 96.11%	96.56
Alaraimi et al. [107] 2021	MEN vs. Glioma vs. PT	Transfer learning with AlexNet	No info shared	AUC = 0.976	94.4
		Transfer learning with VGG16	No info shared	AUC = 0.981	100
		Transfer learning with GoogLeNet	No info shared	AUC = 0.986	98.5
Lo et al. [108] 2019	Grade II vs. Grade III vs. Grade IV	Transfer learning with AlexNet	10-fold CV		97.9
Pei et al. [110] 2020	GBM vs. AST vs. OLI	3D CNN	No info shared		74.9
Gu et al. [29] 2021	1. MEN vs. Glioma vs. PT	Custom CNN model	5-fold CV	SEN = 94.64%, PRE = 94.61%, F1 score = 94.70%	96.39
	2. GBM vs. AST vs. OLI	Custom CNN model	5-fold CV	SEN = 93.66%, PRE = 95.12%, F1 score = 94.05%	97.37
Rajini [115] 2019	MEN vs. Glioma vs. PT	Custom CNN model	5-fold CV		98.16
Anaraki et al. [116] 2019	MEN vs. Glioma vs. PT	Custom CNN model	5-fold CV		94.2
Sajjad et al. [90] 2019	MEN vs. Glioma vs. PT	Transfer learning with VGG19	No info shared	SEN = 88.41%, SPE = 96.12%	94.58
Wahlang et al. [118] 2020	Metastasis vs. Glioma vs. MEN	Lenet	No info shared		48

		AlexNet	No info shared		75
Xiao et al. [87] 2021	MEN vs. Glioma vs. PT	Transfer learning with ResNet50	3-fold, 5-fold, 10-fold CV		98.02
Tandel et al. [22] 2020	Normal vs. LGG vs. HGG	Transfer learning with AlexNet	Multiple CV (K2, K5, K10)	RE = 94.85%, PRE = 94.75%, F1 score = 94.8%	95.97
Ayadi et al. [88] 2021	1. Normal vs. LGG vs. HGG	Custom CNN model	5-fold CV		95
	2. MEN vs. Glioma vs. PT	Custom CNN model	5-fold CV		94.74
4 classes					
Ahammed et al. [68] 2019	Grade I vs. Grade II vs. Grade III vs. Grade IV	VGG19	No info shared	PRE = 94.71%, SEN = 92.72%, SPE = 98.13%, F1 score = 93.71%	98.25
Mohammed et al. [48] 2020	EP vs. MEN vs. MB vs. Normal	Custom CNN model	No info shared	SEN = 96%, PRE = 100%	96
Gilanie et al. [112] 2021	AST-I vs. AST-II vs. AST-III vs. AST-IV	Custom CNN model	No info shared		96.56
Artzi et al. [114] 2021	Normal vs. PA vs. MB vs. EP	Custom CNN model	5-fold CV		88
Rajini [115] 2019	Normal vs. Grade II vs. Grade III vs. Grade IV	Custom CNN model	5-fold CV		96.77
Anaraki et al. [116] 2019	Normal vs. Grade II vs. Grade III vs. Grade IV	Custom CNN model	5-fold CV		

Table 4.2. (*Continued*).

Author & Year	Classification Tasks	Model Architecture	Validation	Performance	ACC%[1]
Sajjad et al. [90] 2019	Grade I vs. Grade II vs. Grade III vs. Grade IV	Transfer learning with VGG19	No info shared		90.67
Xiao et al. [87] 2021	MEN vs. Glioma vs. PT vs. Normal	Transfer learning with ResNet50	3-fold, 5-fold, 10-fold CV	PRE = 97.43%, RE = 97.67%, SPE = 99.24%, F1 score = 97.55%	97.7
Tandel et al. [22] 2020	Normal vs. AST vs. OLI vs. GBM	Transfer learning with AlexNet	Multiple CV (K2, K5, K10)	RE = 94.17%, PRE = 95.41%, F1 score = 94.78%	96.56
Ayadi et al. [88] 2021	1. normal vs. AST vs. OLI vs. GBM	Custom CNN model	5-fold CV		94.41
	2. Grade I vs. Grade II vs. Grade III vs. Grade IV	Custom CNN model	5-fold CV		93.71
5 classes					

Tandel et al. [22] 2020	AST-II vs. AST-III vs. OLI-II vs. OLI-III vs. GBM-IV	Transfer learning with AlexNet	Multiple CV (K2, K5, K10)	RE = 84.4%, PRE = 89.57%, F1 score = 86.89%	87.14
Ayadi et al. [88] 2021	AST-II vs. AST-III vs. OLI-II vs. OLI-III vs. GBM	Custom CNN model	5-fold CV		86.08
6 classes					
Tandel et al. [22] 2020	Normal vs. AST-II vs. AST-III vs. OLI-II vs. OLI-III vs. GBM-IV	Transfer learning with AlexNet	Multiple CV (K2, K5, K10)	RE = 91.51%, PRE = 92.46%, F1 score = 91.97%	93.74
Ayadi et al. [88] 2021	normal vs. AST-II vs. AST-III vs. OLI-II vs. OLI-III vs. GBM	Custom CNN model	5-fold CV		92.09

Many attempts have been made to investigate the value of transfer learning techniques for brain tumor classification [36][42][47][92][96][100][108][113]. Deepak and Ameer [36] used the GoogLeNet with transfer learning technique to differentiate among glioma, MEN, and PT from the dataset provided by Cheng [52]. This proposed system achieved a mean classification accuracy of 98%.

In a study conducted by Yang et al. [42], AlexNet and GoogLeNet were both trained from scratch and fine-tuned from pre-trained models from the ImageNet database to HGG and LGG classification. The dataset used in this method consists of ceT1w images from 113 patients (52 LGG, 61 HGG) with pathologically proven gliomas. The results show that GoogLeNet proved superior to AlexNet for the task. The performance measures, including validation accuracy, test accuracy, and test AUC of GoogLeNet trained from scratch, were 0.867, 0.909, and 0.939, respectively. With fine-tuning, pre-trained GoogLeNet obtained better performance for glioma grading with validation accuracy of 0.867, test accuracy of 0.945, and test AUC 0.968.

Authors in [47] proposed a block-wise fine-tuning strategy using a pre-trained VGG19 for brain tumor classification. The dataset consists of 3064 images (708 MEN, 1426 glioma, and 930 PT) from 233 patients (82 MEN, 89 glioma, and 62 PT). The authors achieved an overall accuracy of 94.82% under five-fold cross-validation. In another study by Bulla et al. [100], the classification was performed in a pre-trained InceptionV3 CNN model with data from the same dataset. Several validation methods, including holdout validation, 10-fold cross-validation, stratified 10-fold cross-validation, and group 10-fold cross-validation, were used during the training process. The best classification accuracy 99.82% of patient-level classification was obtained under group 10-fold cross-validation.

Authors in [96] used InceptionResNetV2, DenseNet121, MobileNet, InceptionV3, Xception, VGG16, and VGG19 which have already been pre-trained on the ImageNet dataset for the classification between HGG and LGG brain images. The MR images used in this research are collected from the BraTS 2019 database, containing 285 patients (210 HGG, 75 LGG). The 3D MRI volumes from the dataset are then converted into 2D slices, generating 26532 images of LGG and 94284 images of HGG. The authors selected 26532 images from HGG to balance these two classes to reduce the impact on classification performance due to class imbalance. The average precision, f1-score, and sensitivity on the test dataset are 98.67%, 98.62%, and 98.33%, respectively.

Lo et al. [108] used transfer learning with fine-tune AlexNet and data augmentation to classify Grade II, Grade III, and Grade IV brain tumor images from a small dataset with 130 patients (30 Grade II, 43 Grade III, 57 Grade IV). The results demonstrate much higher accuracy using pre-trained AlexNet. The proposed transferred DCNN CADx system achieved a mean accuracy of 97.9% and a mean AUC of 0.9991, while the DCNN without pre-trained features only achieved a mean accuracy of 61.42% and a mean AUC of 0.8222.

Kulkarni and Sundari [113] utilized five transfer learning architectures, AlexNet, VGG16, ResNet18, ResNet50, and GoogLeNet, to classify benign and malignant brain tumors from the private dataset collected by the authors, which contains only 200 images (100 benign and 100 malignant). In addition, data augmentation techniques, including scaling, translation, rotation, translation, shearing, and reflection, were performed to generalize the model and reduce the overfitting possibilities. The results show that the fine-tuned AlexNet architecture achieved the highest accuracy and sensitivity of 93.7% and 100%.

Despite many studies on CADx systems that have demonstrated inspiring classification performance, the validating of their algorithms for clinical practice has hardly been carried out. External validation is an efficient approach to overcome the problem caused by data mismatch and to improve the generalization, stability, and robustness of classification algorithms. It is the action of evaluating the classification model in a new independent dataset to determine whether the model performs well. However, we found only two studies that used an external clinical dataset to evaluate the effectiveness and generalization capability of the proposed scheme, described in the following.

Decuyper et al. [69] proposed a 3D CNN model to classify the brain MR volumes collected from TCGA-LGG, TCGA-GBM, and BraTS 2019 databases into HGG and LGG. Multiple MRI sequences, including T_1w, ceT_1w, T_2w, and FLAIR, were used in this research. All MR data were co-registered to the same anatomical template and interpolated to 1 mm^3 voxel sizes. Additionally, a completely independent dataset of 110 patients acquired at the Ghent University Hospital (GUH) was used as an external dataset to validate the efficiency and generalization of the proposed model. The resulting validation accuracy, sensitivity, specificity, and AUC on GUH dataset are 90.00%, 90.16%, 89.80%, and 0.9398.

Gilanie et al. in [112] presented an automatic method using CNN architecture for astrocytoma grading between AST-I, AST-II, AST-III, and AST-IV. The dataset consists of MR slices of 180 subjects, including 50 AST-I cases, 40 AST-II cases, 40 AST-III cases, and 50 AST-IV cases. T1w, T2w, and FLAIR have been used in the experiments. In addition, N4ITK method [119] was used in the preprocessing stage to correct bias field distortion present in MR images. Results have been validated on a locally developed dataset to evaluate the effectiveness and generalization capability of the proposed scheme. The proposed method obtained an overall accuracy of 96.56% on the external validation dataset.

In brain tumor classification, it is often necessary to use image co-registration to preprocess input data when images were collected from different sequences or different scanners. However, we found that this problem has not yet been taken seriously. In the surveyed articles, six studies [69][88][94][110][115,116] used data from multi datasets for one classification target, while only two studies [69][94] performed image co-registration during the image preprocessing process.

Authors in [94] proposed a 2D Mask RCNN model and a 3DConvNet model for automatic and non-invasively distinguishing LGG (Grades II and Grade III) and HGG (Grade IV) on multiple MR sequences of T_1w, ceT_1w, T_2w, and FLAIR. TCIA-LGG and BraTS 2018 databases were used to train and validate these two CNN models in this research work. In the 2D Mask RCNN model, all input MR images were first preprocessed by rigid image registration and intensity inhomogeneity correction. In addition, data augmentation has also been implemented to increase the size and the diversity of the training data. The performance measures, including accuracy, sensitivity, and specificity, were 96.3%, 93.5%, and 97.2% achieved by the proposed 2D Mask RCNN-based method and 97.1%, 94.7%, and 96.8% for the 3DConvNet method, respectively.

In the study conducted by Ayadi [88], the researchers built a custom CNN model for multiple classification tasks. They collected data from three online databases, Radiopaedia, the dataset provided by Cheng, and REMBRANDT for brain tumor classification, while no image co-registration was performed to minimize shifts between images and reduce its impact on the classification performance. The overall accuracy obtained for tumorous and normal classification reaches 100%, for normal, LGG, and HGG classification 95%, for MEN, glioma, and PT classification 94.74%, for normal, AST, OLI, and GBM classification 94.41%, for Grade I, Grade II, Grade III, and Grade IV classification 90.35%, for AST-II, AST-III, OLI-II, OLI-III, and GBM classification 86.08%, and for normal, AST-II, AST-III, OLI-II, OLI-III, and GBM 92.09%.

Authors in [110] proposed a 3D CNN model for brain tumor classification between GBM, AST, and OLI. A merged dataset from CPM-RadPath 2019 and BraTS 2019 databases was used to train and validate the proposed model, while they did not perform image co-registration. The result shows that the classification model has very poor performance for brain tumor classification, with an accuracy of 74.9%.

In [115], the researchers presented a CNN-PSO method for two classification tasks, Normal vs. Grade II vs. Grade III vs. Grade IV and MEN vs. glioma vs. PA. The MR images used for the first task were collected from four publicly available datasets, the IXI dataset, REMBRANDT, TCGA-GBM, and TCGA-LGG. The overall accuracy obtained is 96.77% for classification between normal, Grade II, Grade III, and Grade IV and 98.16% for MEN, glioma, and PA classification.

Similar to the work conducted in [115], Anaraki et al. [116] used MR data merged from four online databases, IXI dataset, REMBRANDT, TCGA-GBM, TCGA-LGG, and one private dataset collected by authors for the classification between normal, Grade II, Grade III, and Grade IV, and the dataset proposed by Cheng [52] for MEN, glioma, and PA classification. Different data augmentation methods were performed to enlarge the size of the dataset, including rotation, translation, scaling, and mirroring. The authors in these studies did not co-register MR images of different sequences from different institutions for the 4-class classification task. The results show the accuracy is 93.1% for normal, Grade II, Grade III, and Grade IV classification and 94.2% for MEN, glioma, and PA classification.

Despite the high accuracy gained in most studies by CNN techniques, we found in several studies [92][109,110][117], the models obtained very poor performance for brain tumor classification.

Authors in [92] explored transfer learning techniques for brain tumor classification. The experiments were performed on the BraTS 2019 dataset which consists of 335 patients diagnosed with brain tumors (259 patients with HGG and 76 patients with LGG). The model achieved a classification AUC of 82.89% on a separate test dataset of 66 patients. The classification performance obtained by transfer learning in this study is relatively low, hindering its following development and application into clinical practice. [109] presented a 3D CNN model developed to categorize adult diffuse glioma cases into OLI and AST classes. The dataset used in the experiment consists of 32 patients (16 patients with OLI and 16 patients with AST). Multiple preprocessing methods were applied, including bias field correction, skull stripping, co-registration, and data augmentation (rotation and cropping). The model achieved accuracy values of 80%. The main reason for the poor performance lies in the small dataset with only 32 patients for model training. It is far from enough to train a 3D model.

In another study [117], two brain tumor classification tasks were studied using Lenet, AlexNet, and U-net CNN architectures. In the experiments, MR images of 11 patients (2 metastasis, 6 glioma, and 3 MEN) from Radiopaedia were utilized to classify metastasis, glioma, and MEN, the data of 20 patients collected from BraTS 2017 were used for HGG and LGG classification. The results show poor classification performance of three CNN architectures on two tasks, with an accuracy of 75% by AlexNet, 48% obtained by Lenet for the first task, and 62% by AlexNet, 60% obtained by U-net for the second task. The poor performance of Lenet is probably due to its simple architecture that is not capable of high-resolution image classification. On the other hand, U-net CNN performs well for segmentation tasks but is not the most used network for classification.

Even though, in the majority of the reviewed studies, CNNs have demonstrated remarkable performances in brain tumor classification, their level of trustworthiness and transparency must be evaluated in a clinic context. Of the included articles, only one study, conducted by Artzi et al. [114], has investigated the Black-Box nature of CNN models for brain tumor classification to ensure that the model is looking at the correct place rather than noise or unrelated artifacts. The authors proposed a pre-trained ResNet-50 CNN architecture to classify three posterior fossa tumors from a private dataset and explained the classification decision by using Gradient-weighted Class Activation Mapping (Grad-CAM). The dataset consists of 158 MRI scans of 22 healthy controls, 63 PA, 57 MB, and 16 EP patients. In this study, several preprocessing methods were used to reduce the influence of MRI data on the classification performance of the proposed CNN model. Image co-registration was performed to ensure that the images become spatially aligned. Bias field correction was also conducted to remove the intensity gradient from the image. Data augmentation methods, including flipping, reflection, rotation, and zooming, were used to increase the size and diversity of the dataset. However, class imbalance within the dataset, particularly the under-representation of EP, was not addressed. The proposed architecture achieved a mean validation accuracy of 88% and 87% for the test dataset. The result demonstrates that the proposed network with Grad-CAM can identify the area of interest and train the classification model based on the pathology-related features.

5. Discussion

Many articles included in this review demonstrate that CNN-based architectures can be powerful and effective when applied to different brain tumor classification tasks. Table 4.2 shows that the classification of images as HGG and LGG, and the differentiation between MEN, glioma, and PT, were the most frequently studied applications. The popularity of these applications is likely linked to the availability of well-known and easily accessible public databases, such as the BraTS datasets and the dataset made available by Cheng [52]. We compared the articles that studied the classification of HGG and LGG and found that the classification performance varies widely even between the articles published in 2021 that utilized state-of-the-art CNN techniques. For example, the study [96], which used transfer learning for brain tumor classification with augmented MRI data from the BraTS 2019 dataset, obtained very high classification performance with a precision of 98.67%, while another study [92] that also used transfer learning technique with advanced CNN-based architecture, AlexNet, for brain tumor classification on MRI data from the same dataset, achieved a quite low classification AUC of 82.89% on an independent test dataset. It is possible that this result was influenced by the lack of data and feature variability, which can be mitigated with data augmentation techniques. In addition, we also observed that one of the key factors that significantly affect the performance of CNN models for brain tumor classification lies in the size of the datasets. Authors in [37] and [95] both proposed custom CNN models to classify HGG and LGG images of 285 MRI scans from the BraTS 2017 dataset. The overall accuracy is 90.7% and 94.28%, respectively. [117] utilized AlexNet for the same task, while the MRI data of only 20 patients from the same dataset were studied. The model in this study yielded poor classification accuracy of 62%, the lowest value among the articles on this classification task.

Among the 61 reviewed articles, 20 articles explored the classification between MEN, glioma, and PT on the MRI data from the same dataset by Cheng [52]. It was demonstrated in Table 4.2 and Figure 6 that all articles achieved relatively high classification accuracy, ranging between 94.2% and 99.82%, the latter being obtained by [100], who applied transfer learning with an InceptionV3 CNN model for the classification task. Among them, five articles [42][47][100][103][107] explored transfer learning with CNN models pre-trained from the ImageNet dataset. [107] utilized transfer learning with the most popular pre-trained CNN architectures VGG16, AlexNet, and GoogLeNet for MEN, glioma, and PT classification. Before feeding the input images into the networks, preprocessing steps, including normalization, resizing, and data augmentation (rotation, cropping, flipping, scaling, and translation) were performed. The result showed that VGG16 attains the best accuracy of 98.92%, higher than pre-trained AlexNet 97.6% and GoogLeNet 98.3%. Another study by [47] did not perform data augmentation before propagating the input data into the pre-trained VGGl9, obtaining a lower accuracy by 4.1% compared with the accuracy obtained by VGG16 in [107], while research in [120] found that the classification error of VGG models decreases with the increased convolution depth from 11 to 19 layers, that is, VGG19 is supposed to have better performance than VGG16. [100] and [103] both used the pre-trained InceptionV3 CNN model for the classification task. The major difference between the two studies lies in the validation methods. Bulla et al. performed various validation methods, including holdout validation, 10-fold cross-validation, stratified 10-fold cross-validation, and group 10-fold cross-validation during the training process, yielding the best accuracy of 99.82% for group 10-fold cross-validation.

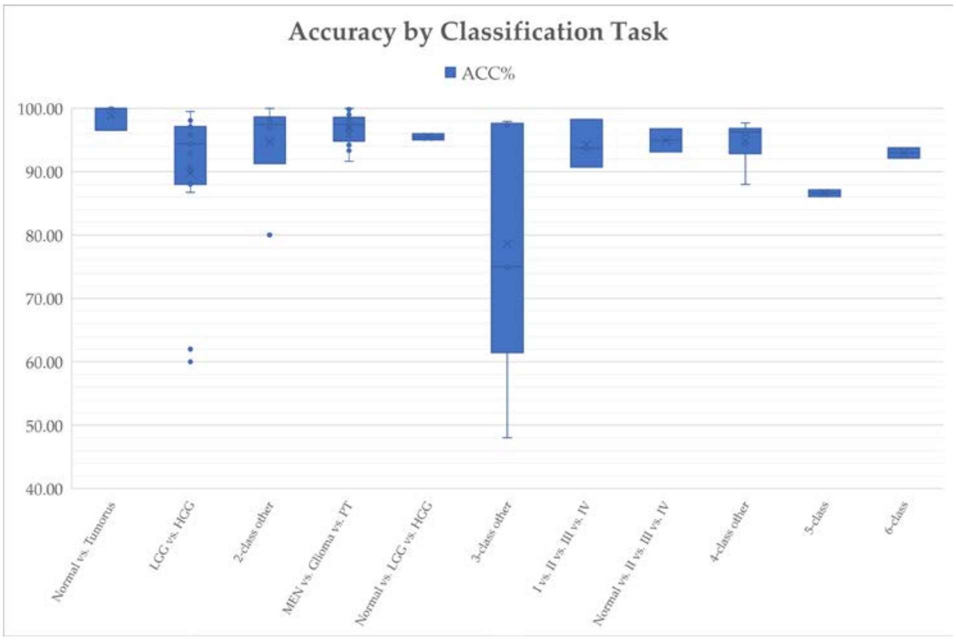


Figure 6. Classification accuracy by classification task.

Figure 6 presents the overall accuracies achieved by the reviewed studies that worked on different classification tasks. What stands out in the figure is that except for the five-class tasks with accuracies lower than 90%, CNNs have achieved promising accuracies on different brain tumor classification tasks, especially the 3-class classification between MEN, glioma, and PT. It is also noticed that the accuracies of 3-class classification tasks fluctuated widely, with the lowest accuracy being 48% in [117] for the metastasis vs. glioma vs. MEN classification. More research attention should be paid to improving the accuracies of these classification tasks.

Figure 7 reveals that there is an increase in the overall accuracy achieved by CNN architectures for brain tumor classification from 2018 to 2021. It is observed that from 2019 the overall classification accuracy achieved in most studies reached 90%, only few works obtained lower accuracies, and the extreme outlier accuracy is 48% in [117] published in 2020. It is also apparent from this figure that the proportion of papers with an accuracy higher than 95% increases from 2020.

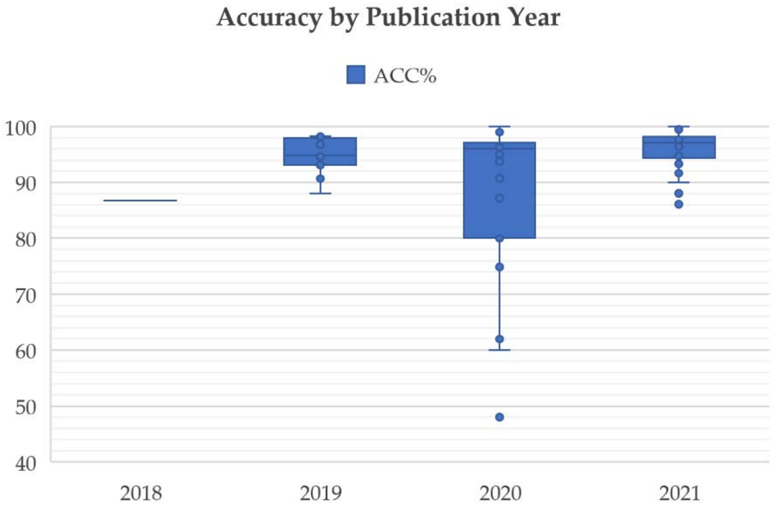


Figure 7. Classification accuracy by publication year.

The overall accuracies by different CNN architectures that were used extensively for brain tumor classification are summarized in Figure 8. It shows that the majority of CNN models have achieved high performance for brain tumor classification tasks, in which transfer learning with ResNet, VGG, and GoogleNet have shown more stable performance than other models, like 3D CNN. Among the reviewed articles, five articles utilized 3D CNN for brain tumor classification and the classification accuracy fluctuates wildly. The highest accuracy was 97.1%, achieved by Zhuge et al. [94], who trained a 3D CNN architecture with a dataset of 315 patients (210 HGG, 105 LGG). The lowest accuracy of 75% was obtained by Pei et al. [110] who used 398 brain MR image volumes for GBM vs. AST vs. OLI classification. In another study [109], authors explored a 3D CNN model for the classification between OLI and AST from a very small dataset of 32 patients (16 OLI, 16 AST) and obtained a low accuracy of 80%. 3D CNN is a promising technique for realizing patient-wise diagnosis, and the accessibility of a large MRI dataset can hopefully improve the performance of 3D CNNs on brain tumor classification.

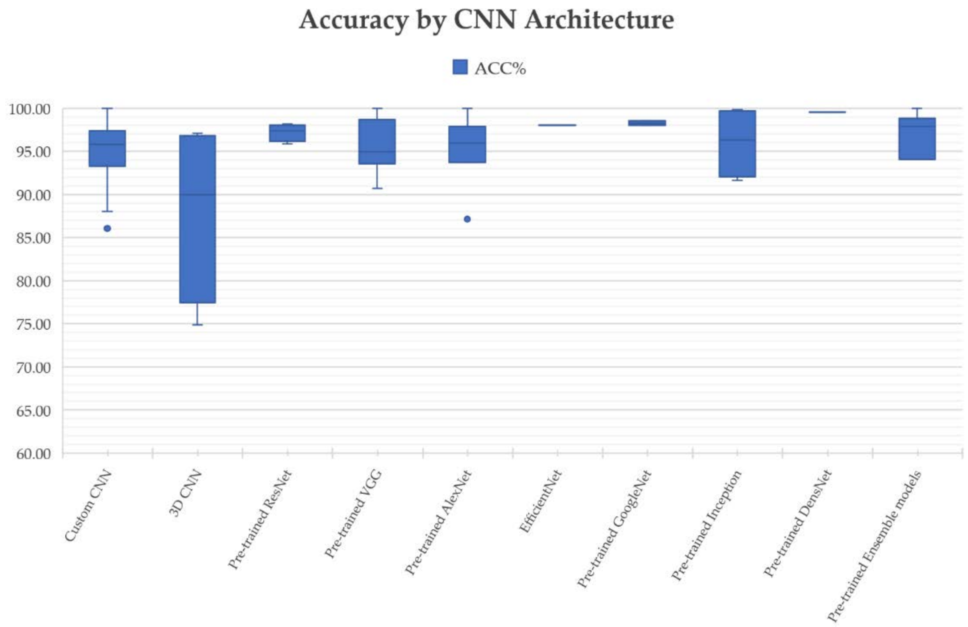


Figure 8. Classification accuracy by CNN architecture.

As was mentioned in the previous chapter, the dataset size is considered a critical factor in determining the classification performance of a CNN architecture. Figure 9 and Figure 10 sum up the classification accuracy obtained with different sizes of datasets, and Figure 9 shows that there is a marked increase in the overall accuracy achieved with more MRI data for brain tumor classification.

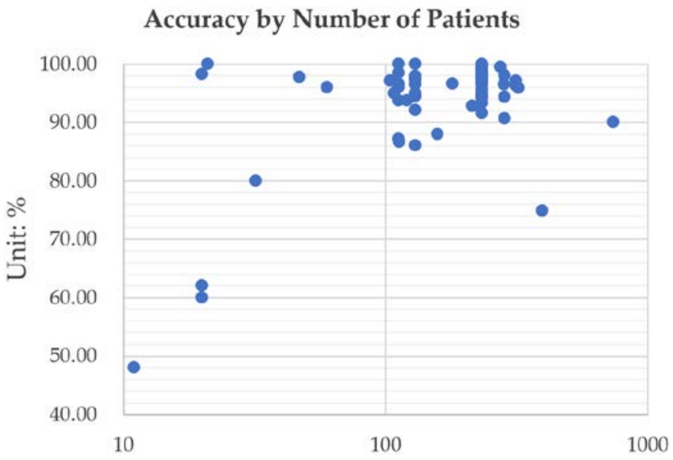


Figure 9. Classification accuracy by number of patients.

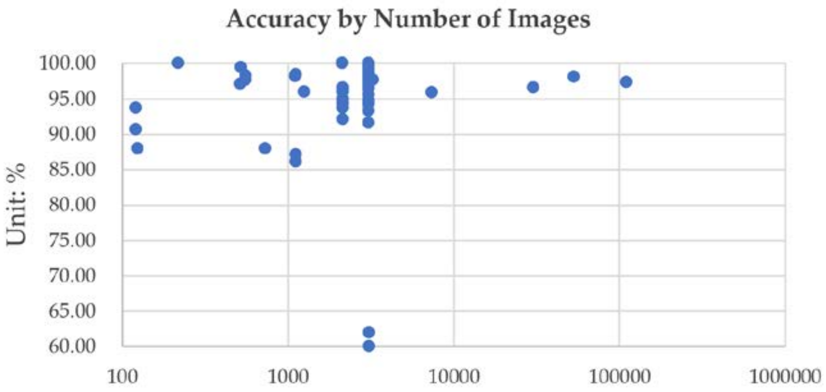


Figure 10. Classification accuracy by number of images.

Researchers have paid increasing attention to enhancing input image quality by doing different preprocessing steps on the brain MRI datasets before propagating into the CNN architectures. Figure 11 presents the overall accuracy obtained with different numbers of preprocessing operations. It shows that studies that pre-processed input MR images collectively obtained higher classification accuracies than studies that performed no preprocessing methods. However, it is not obvious that more steps lead to better performance.

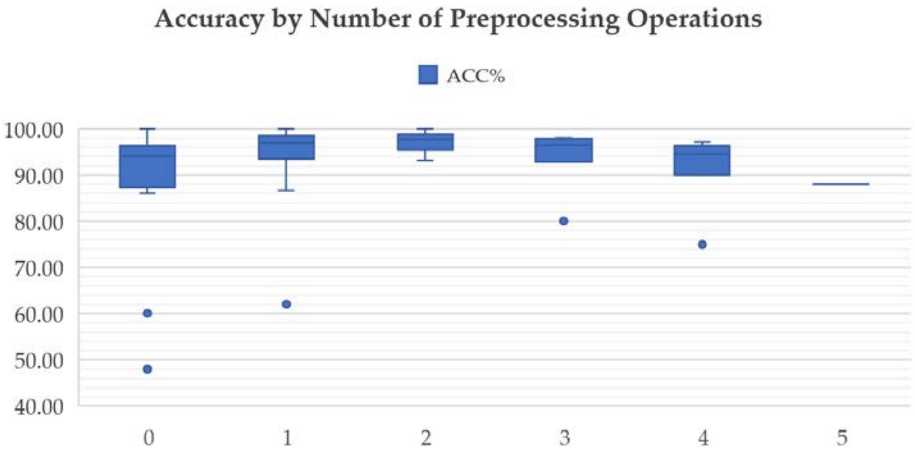


Figure 11. Classification accuracy by number of preprocessing operations.

As previously stated, data augmentation can create variations of the images that can improve the generalization capability of the models to new images, and different data augmentation techniques have been widely explored and applied to increase both the amount and the diversity of the training data. Figure 12 illustrates the overall accuracy obtained with different numbers of data augmentation operations. It can be seen that studies that performed five data augmentation techniques achieved higher and more stable classification performance than studies that performed fewer operations. However, no studies have systematically tested the number and combination of operations that optimise classification accuracy.

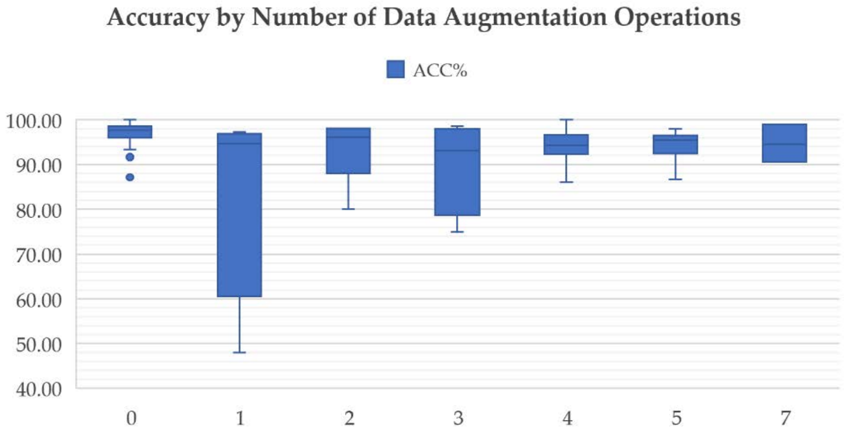


Figure 12. Classification accuracy by number of data augmentation operations.

Beyond showing accuracy gains, the surveyed articles rarely examined their generalization capability and interpretability. Only very few studies [69][112] tested their classification models on an independent dataset, and only one study [114] investigated the Black-Box characteristic of CNN models for brain tumor classification to ensure that the model they obtained was looking at the correct place for decision-making, rather than within the noise or unrelated artifacts.

6. Conclusion

CADx systems may play an important role in assisting physicians in making decisions. This paper surveyed 61 articles that adopted CNNs for brain MRI classification and analyzed the challenges and barriers that CNN-based CADx brain tumor classification systems face today in clinical application and development. The proposed challenges can help advance progress in this field if appropriately addressed.

When considering future directions, despite the achievements of CNNs, we still face challenges in translating and developing them into clinical practice. The Black-Box nature of deep CNNs has greatly limited its application outside a research context. To trust systems powered by CNN models, clinicians need to know how they make predictions. However, among the articles surveyed, very few addressed this problem. Authors in [121] proposed a prototypical part network (ProtoPNet) that can highlight image regions used for decision-making and explain the reasoning process of the classification target by comparing the representative patches of the test image with the prototypes learned from a large amount of data. To date, several studies have tested the explanation model proposed in [121] able to highlight image regions used for decision making in medical imaging fields, such as mass lesion classification [122], lung disease detection [123,124], and Alzheimer's diseases classification [125]. Those developments suggest directions for future research in the brain tumor classification field to tame the Black-Box problem.

With a limited number of training data, transfer learning with fine-tuning on pre-trained CNNs was demonstrated to yield better results for brain tumor classification than training such CNNs from scratch [42][108]. It is efficient for training networks when training data is expensive or difficult to collect in medical fields. In addition, high hardware requirements and long training time are also the challenges that CNN-based CADx brain tumor classification systems face today in clinical applications. The continued development of state-of-the-art CNN architectures has come with a voracious appetite for computing power. Since the cost of training a deep learning model scales with the number of parameters and the amount of input data, this implies that computational requirements grow as at least the square of the number of training data [126]. With pre-trained models, transfer learning is also promising to address the difficulties caused by high hardware requirements and long training time when adopting CNN-based CADx systems for brain tumor classification in clinical practice.

In light of the limitations and challenges mentioned above, there are still some general limitations that hinder the clinic adoption of CNN-based CADx systems. CADx systems are mainly used for educational and training purposes but not in clinical practice. The majority of clinics still hesitate to use CADx-based systems. One reason for this is the lack of standardized methods for evaluating CADx systems. Another reason lies in some technical weaknesses involved in the task of making diagnostic decisions with CADx systems. It is the lack of training of physicians on how to interact with the system and how to interpret the results, which can be a key reason for the poor performance of CADx systems. Therefore, proper education and training for physicians on how to use the systems and interpret the outcomes of CAD systems are important.

In short, the future of CNN-based brain tumor classification studies is very promising and focusing on the right direction with references to the challenges mentioned above would advance these studies from research labs to hospitals. We believe that our review provides researchers in the biomedical and machine learning communities with indicators for useful future directions for this purpose.

Author Contributions: Conceptualization, C.T., D.N.M., F.Z., L.R., Y.T.X.; Methodology, C.T., D.N.M., F.Z., L.R., Y.T.X.; Writing—original draft preparation, Y.T.X.; Writing—review and editing, Ca.T., C.T., D.N.M, F.Z., L.R.; Resources, Ca.T., R.A., R.L.; Supervision, Ca.T., C.T., D.N.M; All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by the China Scholarship Council (No. 202008320283).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

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

References

1. Louis, D. N.; Perry, A.; Wesseling, P.; Brat, D. J.; Cree, I. A.; et al. The 2021 WHO classification of tumors of the central nervous system: a summary. *Neuro-oncology* **2021**, *23*, 1231-1251. [CrossRef]
2. Cancer Research UK. https://www.cancerresearchuk.org/health-professional/cancer-statistics/statistics-by-cancer-type/brain-other-cns-and-intracranial-tumours/incidence%23collapseTen#heading-One (archived on 10 February 2022).
3. Islami F; Ward E M; Sung H; et al. Annual report to the nation on the status of cancer, part 1: national cancer statistics[J]. *JNCI: Journal of the National Cancer Institute* **2021**, *113*, 1648-1669. [CrossRef]
4. Johnson DR; O'Neill BP. Glioblastoma survival in the United States before and during the temozolomide era. *J Neurooncol.* **2012**, *107*, 359-64. [CrossRef]
5. Gao H, Jiang X. Progress on the diagnosis and evaluation of brain tumors[J]. *Cancer Imaging* **2013**, *13*, 466. [CrossRef]
6. Villanueva-Meyer, J. E.; Mabray, M. C.; Cha, S. Current clinical brain tumor imaging. Neurosurgery 2017, 81, 397-415. [CrossRef] [PubMed]
7. Zaccagna, F.; Riemer, F.; Priest, A. N.; McLean, M. A.; et al. A. Non-invasive assessment of glioma microstructure using VERDICT MRI: correlation with histology. *European radiology* **2019**, *29*, 5559-5566. [CrossRef]
8. Radbruch A; Wiestler B; Kramp L; et al. Differentiation of glioblastoma and primary CNS lymphomas using susceptibility weighted imaging. *Eur J Radiol.* **2013**, *82*, 552-556. [CrossRef]

9. Xiao H-F; Chen Z-Y; Lou X; et al. Astrocytic tumour grading: a comparative study of three-dimensional pseudo continuous arterial spin labelling, dynamic susceptibility contrast-enhanced perfusion-weighted imaging, and diffusion- weighted imaging. *Eur Radiol.* **2015**, *25*, 3423-3430. [CrossRef]
10. Zaccagna, F.; Grist, J. T.; Quartuccio, N.; Riemer, F.; Fraioli, F.; et al. Imaging and treatment of brain tumors through molecular targeting: Recent clinical advances. *European Journal of Radiology* **2021**, *142*, 109842. [CrossRef]
11. Figueiredo P; Figueiredo I; Pinto L; Kumar S; Tsai Y; Mamonov A. Polyp detection with computer-aided diagnosis in white light colonoscopy: Comparison of three different methods. *Endosc Int Open* **2019,** *7*, E209-E215. [CrossRef]
12. Yeung M.; Sala E.; Schönlieb C.B.; Rundo L. Focus U-Net: A novel dual attention-gated CNN for polyp segmentation during colonoscopy. *Computers in Biology and Medicine* **2021**, *137*, 104815. [CrossRef]
13. Gong, J.; Liu, J. Y.; Sun, X. W.; Zheng, B.; Nie, S. D. Computer-aided diagnosis of lung cancer: the effect of training data sets on classification accuracy of lung nodules. *Physics in Medicine & Biology* **2018**, *63*, 035036. [CrossRef] [PubMed]
14. Nishio M; Sugiyama O; Yakami M; Ueno S; Kubo T; Kuroda T; Togashi K. Computer-aided diagnosis of lung nodule classification between benign nodule, primary lung cancer, and metastatic lung cancer at different image size using deep convolutional neural network with transfer learning. *PLoS One* **2018**, *13*, e0200721. [CrossRef]
15. Buchlak Q D; Esmaili N; Leveque J C; et al. Machine learning applications to neuroimaging for glioma detection and classification: An artificial intelligence augmented systematic review[J]. *Journal of Clinical Neuroscience* **2021**, *89*, 177-198. [CrossRef]
16. Ahmadi M; Dashti Ahangar F; Astaraki N; et al. FWNNet: Presentation of a New Classifier of Brain Tumor Diagnosis Based on Fuzzy Logic and the Wavelet-Based Neural Network Using Machine-Learning Methods[J]. *Computational Intelligence and Neuroscience* **2021**, 2021. [CrossRef]
17. Sengupta A; Ramaniharan A K; Gupta R K; et al. Glioma grading using a machine-learning framework based on optimized features obtained from T1 perfusion MRI and volumes of tumor components[J]. *Journal of Magnetic Resonance Imaging* **2019**, *50*, 1295-1306. [CrossRef]
18. Hu J; Wu W; Zhu B; et al. Cerebral glioma grading using Bayesian network with features extracted from multiple modalities of magnetic resonance imaging[J]. *PLoS One* **2016**, *11*, e0153369. [CrossRef] [PubMed]
19. Raju A R; Suresh P; Rao R R. Bayesian HCS-based multi-SVNN: a classification approach for brain tumor segmentation and classification using Bayesian fuzzy clustering[J]. *Biocybernetics and Biomedical Engineering* **2018**, 38, 646-660. [CrossRef]
20. Schulz M A; Yeo B T; Vogelstein J T; et al. Different scaling of linear models and deep learning in UKBiobank brain images versus machine-learning datasets[J]. *Nature communications* **2020**, *11*, 1-15. [CrossRef]
21. Schmidhuber J. Deep learning in neural networks: an overview. *Neural Netw* **2015**, *61*, 85–117. [CrossRef]
22. Tandel G S; Balestrieri A; Jujaray T; et al. Multiclass magnetic resonance imaging brain tumor classification using artificial intelligence paradigm[J]. *Computers in Biology and Medicine* **2020**, *122*, 103804. [CrossRef]
23. Shen D; Wu G; Suk H I. Deep learning in medical image analysis[J]. *Annual review of biomedical engineering* **2017**, *19*, 221-248. [CrossRef]
24. Yasaka K; Akai H; Kunimatsu A; et al. Deep learning with convolutional neural network in radiology[J]. *Japanese journal of radiology* **2018**, *36*, 257-272. [CrossRef]
25. Özcan H; Emiroğlu B G; Sabuncuoğlu H; et al. A comparative study for glioma classification using deep convolutional neural networks. [J]. *Mathematical Biosciences and Engineering: MBE* **2021**, *18*, 1550-1572. [CrossRef] [PubMed]
26. Díaz-Pernas, F.J.; Martínez-Zarzuela, M.; Antón-Rodríguez, M.; González-Ortega, D. A. Deep Learning Approach for Brain Tumor Classification and Segmentation Using a Multiscale Convolutional Neural Network. *Healthcare* **2021**, *9*, 153. [CrossRef]
27. Mzoughi H; Njeh I; Wali A; et al. Deep multi-scale 3D convolutional neural network (CNN) for MRI gliomas brain tumor classification[J]. *Journal of Digital Imaging* **2020**, *33*, 903-915. [CrossRef]
28. Abd El Kader I; Xu G; Shuai Z; et al. Differential deep convolutional neural network model for brain tumor classification[J]. *Brain Sciences* **2021**, *11*, 352. [CrossRef] [PubMed]
29. Gu X; Shen Z; Xue J; et al. Brain Tumor MR Image Classification Using Convolutional Dictionary Learning With Local Constraint[J]. *Frontiers in Neuroscience* **2021**, *15*. [CrossRef] [PubMed]
30. Avorn J; Fischer M. 'Bench to behavior': translating comparative effectiveness research into improved clinical practice[J]. *Health Affairs* **2010** *29*, 1891-1900. [CrossRef] [PubMed]
31. Zadeh Shirazi A; Fornaciari E; McDonnell M D; et al. The application of deep convolutional neural networks to brain cancer images: a survey[J]. *Journal of Personalized Medicine* **2020**, *10*, 224. [CrossRef]
32. Nazir M; Shakil S; Khurshid K. Role of Deep Learning in Brain Tumor Detection and Classification (2015 to 2020): A Review[J]. *Computerized Medical Imaging and Graphics*, **2021**, 101940. [CrossRef]
33. Muhammad K; Khan S; Del Ser J; et al. Deep learning for multigrade brain tumor classification in smart healthcare systems: A prospective survey[J]. *IEEE Transactions on Neural Networks and Learning Systems*, **2020**, *32*, 507-522. [CrossRef]
34. Moher D; Liberati A; Tetzlaff J; Altman DG. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *Annals of internal medicine* **2009**, *151*, 264–9. [CrossRef]
35. Miotto R; Wang F; Wang S; et al. Deep learning for healthcare: review, opportunities and challenges[J]. *Briefings in bioinformatics* **2018**, *19*, 1236-1246. [CrossRef]
36. Deepak, S.; P. M. Ameer. Brain tumor classification using deep CNN features via transfer learning. *Computers in biology and medicine 111*, **2019**, 103345. [CrossRef]

37. Ge, C.; Gu, I. Y. H.; Jakola, A. S.; Yang, J. Deep semi-supervised learning for brain tumor classification. *BMC Medical Imaging* **2020**, *20*, 1-11. [CrossRef]
38. Huang, Z.; Xu, H.; Su, S.; Wang, T.; Luo, Y.; Zhao, X.; et al. A computer-aided diagnosis system for brain magnetic resonance imaging images using a novel differential feature neural network. *Computers in biology and medicine* **2020**, *121*, 103818. [CrossRef]
39. Díaz-Pernas F J; Martínez-Zarzuela M; Antón-Rodríguez M; et al. A deep learning approach for brain tumor classification and segmentation using a multiscale convolutional neural network. *Healthcare* **2021**, *9*, 153. [CrossRef]
40. Guan, Y.; Aamir, M.; Rahman, Z. Ali, A.; Abro, W. A.; et al. A framework for efficient brain tumor classification using MRI images. *Mathematical Biosciences and Engineering* **2021**, *18*, 5790-5815. [CrossRef]
41. Gab Allah; Ahmed M.; Amany M. Sarhan; Nada M. Elshennawy. Classification of Brain MRI Tumor Images Based on Deep Learning PGGAN Augmentation. *Diagnostics* **2021**, *11*, 2343. [CrossRef]
42. Yang Y; Yan L F; Zhang X; et al. Glioma grading on conventional MR images: a deep learning study with transfer learning[J]. *Frontiers in neuroscience* **2018**, 804. [CrossRef]
43. Brownlee, Jason. "What is the Difference Between Test and Validation Datasets?". https://machinelearningmastery.com/ difference-test-validation-datasets (Retrieved on 18 February 2022).
44. Prechelt, Lutz; Geneviève B. Orr. "Early Stopping — But When?". In Grégoire Montavon; Klaus-Robert Müller (eds.). Neural Networks: Tricks of the Trade. Lecture Notes in Computer Science. *Springer Berlin Heidelberg* **2012**, 53–67. [CrossRef]
45. F-score, Wikipedia. Available online: https://en.wikipedia.org/wiki/F-score (archived on 22 March 2022)
46. Ismael S A A; Mohammed A; Hefny H. An enhanced deep learning approach for brain cancer MRI images classification using residual networks[J]. *Artificial intelligence in medicine* **2020**, *102*, 101779. [CrossRef]
47. Swati Z N K; Zhao Q; Kabir M; et al. Brain tumor classification for MR images using transfer learning and fine-tuning[J]. *Computerized Medical Imaging and Graphics* **2019**, *75*, 34-46. [CrossRef]
48. Mohammed B A; Al-Ani M S. An efficient approach to diagnose brain tumors through deep CNN[J]. *Math. Biosci. Eng*, **2020**, *18*, 851-867. [CrossRef]
49. Andri Signorell; Ken Aho; Andreas Alfons; Nanina Anderegg; et al. DescTools: Tools for descriptive statistics. R package version 0.99.44. Available online: https://cran.r-project.org/package=DescTools. (accessed on 4 May 2022)
50. The Cancer Genome Atlas, TCGA-GBM. Available online: https://wiki.cancerimagingarchive.net/display/Public/TCGA-GBM.
51. The Cancer Genome Atlas, TCGA-LGG. Available online: https://wiki.cancerimagingarchive.net/display/Public/TCGA-LGG.
52. Figshare, Brain tumor dataset. Available online: https://figshare.com/articles/dataset/brain_tumor_dataset/ 1512427/5.
53. C. Navoneel. Available online: https://www.kaggle.com/navoneel/brain-mri-images-for-brain-tumor-detect.
54. REMBRANDT. Available online: https://wiki.cancerimagingarchive.net/display/Public/REMBRANDT.
55. Brain Tumor Segmentation (BraTS) Challenge. Available online: http://www.braintumorsegmentation.org/.
56. ClinicalTrials.gov. Available online: https://www.clinicaltrials.gov/.
57. Computational Precision Medicine: Radiology-Pathology Challenge on Brain Tumor Classification 2019. Available online: https://www.med.upenn.edu/cbica/cpm-rad-path-2019/.
58. IXI dataset. Available online: https://brain-development.org/ixi-dataset/
59. Rider neuro MRI. Available online: https://wiki.cancerimagingarchive.net/display/Public/RIDER+NEURO+MRI.
60. Harvard Medical School Data. Available online: http://www.med.harvard.edu/AANLIB/.
61. MRI sequence. Wikipedia. Available online: https://en.wikipedia.org/wiki/MRI_sequence (accessed on 18 February 2022).
62. My-MS.org. MRI Basics. Available online: https://my-ms.org/mri_basics.htm (accessed on 18 February 2022).
63. Basic proton MR imaging. Harvard Medical School. Available online: http://www.med.harvard.edu/aanlib/basicsmr.html (accessed on 19 February 2022).
64. Fluid attenuation inversion recovery. Radiopaedia.org. Available online: https://radiopaedia.org/articles/fluid-attenuated-inversion-recovery (accessed on 19 February 2022).
65. Chen M W; King N K K; Selvarajan S; et al. Benign scalp lump as an unusual presentation of extranodal Rosai-Dorfman disease[J]. *Surgical neurology international* **2014**, *5*. [CrossRef]
66. Mohan G; Subashini M M. MRI based medical image analysis: Survey on brain tumor grade classification[J]. *Biomedical Signal Processing and Control* **2018**, *39*, 139-161. [CrossRef]
67. Collewet G; Strzelecki M; Mariette F. Influence of MRI acquisition protocols and image intensity normalization methods on texture classification[J]. *Magnetic resonance imaging* **2004**, 22, 81-91. [CrossRef]
68. KV A M; Rajendran V R. Glioma tumor grade identification using artificial intelligent techniques[J]. *Journal of medical systems* **2019**, *43*, 1-12. [CrossRef]
69. Decuyper M; Bonte S; Deblaere K; et al. Automated MRI based pipeline for segmentation and prediction of grade, IDH mutation and 1p19q co-deletion in glioma[J]. *Computerized Medical Imaging and Graphics* **2021**, 88, 101831. [CrossRef]
70. Hashemi M. Enlarging smaller images before inputting into convolutional neural network: zero-padding vs. interpolation[J]. *Journal of Big Data* **2019**, 6, 1-13. [CrossRef]
71. Hashemi M. Web page classification: a survey of perspectives, gaps, and future directions. *Multimed Tools Appl* **2019**. [CrossRef]
72. Karthick S; Maniraj S. Different medical image registration techniques: a comparative analysis[J]. *Current Medical Imaging* **2019**, *15*, 911-921. [CrossRef]
73. Song S; Zheng Y; He Y. A review of methods for bias correction in medical images[J]. *Biomedical Engineering Review* **2017**, 1. [CrossRef]

74. Introduction to Data Mismatch, Overfitting and Underfitting in Building Machine Learning Systems. Towards Data Science. Available online: https://towardsdatascience.com/introduction-to-overfitting-underfitting-and-data-mismatch-in-building-machine-learning-systems-52f1225a8a35 (accessed on 19 February 2022).
75. Krizhevsky A; Sutskever I; Hinton G E. Imagenet classification with deep convolutional neural networks[J]. *Advances in neural information processing systems* **2012**, *25*. [CrossRef]
76. Simonyan K; Zisserman A. Very deep CNN for large-scale image recognition[J]. **2015**. [CrossRef]
77. Szegedy C; Liu W; Jia Y; et al. Going deeper with convolutions[C]//*Proceedings of the IEEE conference on computer vision and pattern recognition* **2015**, 1-9. [CrossRef]
78. He K; Zhang X; Ren S; et al. Deep residual learning for image recognition[C]//*Proceedings of the IEEE conference on computer vision and pattern recognition* **2016**, 770-778. [CrossRef]
79. Huang G; Liu Z; Van Der Maaten L; et al. Densely connected convolutional networks[C]//*Proceedings of the IEEE conference on computer vision and pattern recognition* **2017**, 4700-4708. [CrossRef]
80. Tan M; Le Q. Efficientnet: Rethinking model scaling for convolutional neural networks[C]//*International conference on machine learning. PMLR* **2019**, 6105-6114. [CrossRef]
81. Srivastava N; Hinton G; Krizhevsky A; et al. Dropout: a simple way to prevent neural networks from overfitting[J]. *The journal of machine learning research* **2014**, *15*, 1929-1958.
82. The Perelman School of Medicine at the University of Pennsylvania. 'Multimodal Brain Tumor Segmentation Challenge 2019'. Available online: http://braintumorsegmentation.org/(accessed on 19 February 2022).
83. Li D C; Liu C W; Hu S C. A learning method for the class imbalance problem with medical data sets[J]. *Computers in biology and medicine* **2010**, *40*, 509-518. [CrossRef]
84. El Kader I A; Xu G; Shuai Z; et al. Brain tumor detection and classification by hybrid CNN-DWA model using MR images[J]. Current Medical Imaging 2021, 17, 1248-1255. [CrossRef]
85. Khan H A; Jue W; Mushtaq M; et al. Brain tumor classification in MRI image using convolutional neural network[J]. *Math. Biosci. Eng* **2020**, *17*, 6203-6216. [CrossRef]
86. Naser M A; Deen M J. Brain tumor segmentation and grading of lower-grade glioma using deep learning in MRI images[J]. *Computers in biology and medicine* **2020**, *121*, 103758. [CrossRef]
87. Xiao G; Wang H; Shen J; et al. Synergy Factorized Bilinear Network with a Dual Suppression Strategy for Brain Tumor Classification in MRI[J]. Micromachines 2022, 13, 15. [CrossRef]
88. Ayadi W; Elhamzi W; Charfi I; et al. Deep CNN for brain tumor classification[J]. *Neural Processing Letters* **2021**, *53*, 671-700. [CrossRef]
89. Amann J; Blasimme A; Vayena E; et al. Explainability for artificial intelligence in healthcare: a multidisciplinary perspective[J]. *BMC Medical Informatics and Decision Making* **2020**, *20*, 1-9. [CrossRef]
90. Sajjad M; Khan S; Muhammad K; et al. Multi-grade brain tumor classification using deep CNN with extensive data augmentation[J]. *Journal of computational science* **2019**, *30*, 174-182. [CrossRef]
91. Weiss K; Khoshgoftaar T M; Wang D D. A survey of transfer learning[J]. *Journal of Big data* **2016**, *3*, 1-40.[CrossRef]
92. Hao R; Namdar K; Liu L; et al. A transfer learning–based active learning framework for brain tumor classification[J]. *Frontiers in Artificial Intelligence* **2021**, *4*. [CrossRef]
93. Tripathi P C; Bag S. A computer-aided grading of glioma tumor using deep residual networks fusion[J]. *Computer Methods and Programs in Biomedicine* **2022**, *215*, 106597. [CrossRef]
94. Zhuge Y; Ning H; Mathen P; et al. Automated glioma grading on conventional MRI images using deep convolutional neural networks[J]. *Medical physics* **2020**, *47*, 3044-3053. [CrossRef]
95. He M; Han K; Zhang Y; et al. Hierarchical-order multimodal interaction fusion network for grading gliomas[J]. *Physics in Medicine & Biology* **2021**, *66*, 215016. [CrossRef]
96. El Hamdaoui H.; Benfares, A.; Boujraf, S.; et al. High precision brain tumor classification model based on deep transfer learning and stacking concepts. *Indonesian Journal of Electrical Engineering and Computer Science* **2021**, *24*, 167-177. [CrossRef]
97. Chikhalikar A M; Dharwadkar N V. Model for Enhancement and Segmentation of Magnetic Resonance Images for Brain Tumor Classification[J]. *Pattern Recognition and Image Analysis* **2021**, *31*, 49-59. [CrossRef]
98. Ahmad, F. Classification on magnetic resonance imaging (Mri) brain tumour using BPNN, SVM and CNN[J]. *International Journal of Recent Technology and Engineering (IJRTE)* **2019**, *8*, 8601-8607. [CrossRef]
99. Alhassan A M; Zainon W M N W. Brain tumor classification in magnetic resonance image using hard swish-based RELU activation function-convolutional neural network[J]. *Neural Computing and Applications* **2021**, *33*, 9075-9087. [CrossRef]
100. Bulla P; Anantha L; Peram S. Deep Neural Networks with Transfer Learning Model for Brain Tumors Classification[J]. *Traitement du Signal* **2020**, *37*. [CrossRef]
101. Ghassemi N; Shoeibi A; Rouhani M. Deep neural network with generative adversarial networks pre-training for brain tumor classification based on MR images[J]. *Biomedical Signal Processing and Control* **2020**, *57*, 101678. [CrossRef]
102. Kakarla J; Isunuri B V; Doppalapudi K S; et al. Three-class classification of brain magnetic resonance images using average-pooling convolutional neural network[J]. *International Journal of Imaging Systems and Technology* **2021**, *31*, 1731-1740. [CrossRef]
103. Noreen N; Palaniappan S; Qayyum A; et al. Brain Tumor Classification Based on Fine-Tuned Models and the Ensemble Method[J]. *CMC-COMPUTERS MATERIALS & CONTINUA* **2021**, *67*, 3967-3982. [CrossRef]

104. Noreen N; Palaniappan S; Qayyum A; et al. A deep learning model based on concatenation approach for the diagnosis of brain tumor[J]. *IEEE Access* **2020**, *8*, 55135-55144. [CrossRef]
105. Kumar R L; Kakarla J; Isunuri B V; et al. Multi-class brain tumor classification using residual network and global average pooling[J]. *Multimedia Tools and Applications* **2021**, *80*, 13429-13438. [CrossRef]
106. Badža M M; Barjaktarović M Č. Classification of brain tumors from MRI images using a convolutional neural network[J]. *Applied Sciences* **2020**, *10*, 1999. [CrossRef]
107. Alaraimi S; Okedu K E; Tianfield H; et al. Transfer learning networks with skip connections for classification of brain tumors[J]. *International Journal of Imaging Systems and Technology* **2021**, *31*, 1564-1582. [CrossRef]
108. Lo C M; Chen Y C; Weng R C; et al. Intelligent glioma grading based on deep transfer learning of MRI radiomic features[J]. *Applied Sciences*, **2019**, *9*, 4926. [CrossRef]
109. Kurc T; Bakas S; Ren X; et al. Segmentation and classification in digital pathology for glioma research: challenges and deep learning approaches[J]. *Frontiers in neuroscience* **2020**, *27*. [CrossRef]
110. Pei L; Vidyaratne L; Rahman M M; et al. Context aware deep learning for brain tumor segmentation, subtype classification, and survival prediction using radiology images[J]. *Scientific Reports* **2020**, *10*, 1-11. [CrossRef]
111. McAvoy M; Prieto P C; Kaczmarzyk J R; et al. Classification of glioblastoma versus primary central nervous system lymphoma using convolutional neural networks[J]. *Scientific Reports* **2021**, *11*, 1-7. [CrossRef]
112. Gilanie G; Bajwa U I; Waraich M M; et al. Risk-free WHO grading of astrocytoma using convolutional neural networks from MRI images[J]. *Multimedia Tools and Applications* **2021**, *80*, 4295-4306. [CrossRef]
113. KULKARNI S M; SUNDARI G. COMPARATIVE ANALYSIS OF PERFORMANCE OF DEEP CNN BASED FRAMEWORK FOR BRAIN MRI CLASSIFICATION USING TRANSFER LEARNING[J]. *Journal of Engineering Science and Technology* **2021**, *16*, 2901-2917.
114. Artzi M; Redmard E; Tzemach O; et al. Classification of pediatric posterior fossa tumors using convolutional neural network and tabular data[J]. *IEEE Access* **2021**, *9*, 91966-91973. [CrossRef]
115. Rajini N H. Brain tumor image classification and grading using convolutional neural network and particle swarm optimization algorithm[J]. *International Journal of Engineering and Advanced Technology (IJEAT)* **2019**, *8*, 2249-8958.
116. Anaraki A K; Ayati M; Kazemi F. Magnetic resonance imaging-based brain tumor grades classification and grading via convolutional neural networks and genetic algorithms[J]. *biocybernetics and biomedical engineering* **2019**, *39*, 63-74. [CrossRef]
117. Wahlang I; Sharma P; Sanyal S; et al. Deep learning techniques for classification of brain MRI[J]. International Journal of *Intelligent Systems Technologies and Applications* **2020**, *19*, 571-588. [CrossRef]
118. Tandel G S; Tiwari A; Kakde O G. Performance optimisation of deep learning models using majority voting algorithm for brain tumour classification[J]. *Computers in Biology and Medicine* **2021**, *135*, 104564. [CrossRef]
119. Tustison NJ; Avants BB; Cook PA; Zheng Y; Egan A; Yushkevich PA; Gee JC. N4ITK: improved N3 bias correction. *IEEE Trans Med Imaging* **2010**, 29, 1310–1320. [CrossRef]
120. Simonyan K; Zisserman A. Very deep CNN for large-scale image recognition[J]. **2015**. [CrossRef]
121. Chen C; Li O; Tao D; et al. This looks like that: deep learning for interpretable image recognition[J]. *Advances in neural information processing systems* **2019**, *32*. [CrossRef]
122. Barnett A J; Schwartz F R; Tao C; et al. A case-based interpretable deep learning model for classification of mass lesions in digital mammography[J]. *Nature Machine Intelligence* **2021**, *3*, 1061-1070. [CrossRef]
123. Singh G; Yow K C. An interpretable deep learning model for COVID-19 detection with chest x-ray images[J]. *IEEE Access* **2021**, *9*, 85198-85208. [CrossRef]
124. Kim E; Kim S; Seo M; et al. XProtoNet: diagnosis in chest radiography with global and local explanations[C]//*Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* **2021**, 15719-15728. [CrossRef]
125. Mohammadjafari S; Cevik M; Thanabalasingam M; et al. Using ProtoPNet for interpretable Alzheimer's disease classification[C]//*Proceedings of the Canadian Conference on Artificial Intelligence* **2021**. [CrossRef]
126. Thompson N C; Greenewald K; Lee K; et al. The Computational Limits of Deep Learning[J]. **2020**. [CrossRef]