A multi-sequences MRI deep framework study applied to glioma classification

Matthieu Coupet¹,² · Thierry Urruty¹,² · Teerapong Leelanupab³ · Mathieu Naudin²,⁴ · Pascal Bourdon²,⁴ · Christine Fernandez Maloigne²,⁴ · Rémy Guillevin²,⁴,⁵

Received: 31 July 2020 / Revised: 2 September 2021 / Accepted: 17 January 2022
© The Author(s), under exclusive licence to Springer Science+Business Media, LLC, part of Springer Nature 2022

Abstract
Glioma is one of the most important central nervous system tumors, ranked 15th in the most common cancer for men and women. Magnetic Resonance Imaging (MRI) represents a common tool for medical experts to the diagnosis of glioma. A set of multi-sequences from an MRI is selected according to the severity of the pathology. Our proposed approach aims more to create a computer-aided system that is capable of helping more the expert diagnose the brain gliomas. more We propose a supervised learning regime based on a convolutional neural network based framework and transfer learning techniques. Our research more focuses on the performance of different pre-trained deep learning models with respect to different MRI sequences. We highlight the best combinations of such model-MRI sequence couple for our specific task of classifying healthy brain against brain with glioma. more We also propose to visually analyze the extracted deep features for studying the existing relation of the MRI sequences and models. This interpretability analysis gives some hints for medical expert to understand the diagnosis made by the models. Our study is based on the well-known BraTS datasets including multi-sequence images and expert diagnosis.

For the Alzheimer’s Disease Neuroimaging Initiative
Data used in preparation of this article were obtained from the Alzheimer’s Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report. A complete listing of ADNI investigators can be found at: http://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf

Teerapong.Leelanupab@it.kmitl.ac.th

¹ XLIM Laboratory, University of Poitiers, UMR CNRS 7252, Poitiers, France
² I3M, Common Laboratory CNRS-Siemens, University and Hospital of Poitiers, Poitiers, France
³ Faculty of Information Technology, King Mongkut’s Institute of Technology Ladkrabang (KMITL), Bangkok, 10520, Thailand
⁴ Poitiers University Hospital, CHU, Poitiers, France
⁵ DACTIM-MIS/LMA Laboratory University of Poitiers, UMR CNRS 7348, Poitiers, France
1 Introduction

According to “Our World in Data” [50], an estimated 10 million people succumbed to cancer in 2017. This reality induces cancer to be the second preeminent cause of death after cardiovascular diseases. Therefore, fighting cancer is a worldwide priority in health.

Our research mainly focuses on gliomas, which are pathologies of the glial cells protecting and supporting the neurons. Those glia cells can be classified according to the cell type, grade, and localization [34]. Gliomas consist of about 30% of all brain tumors, and 80% of the gliomas are malignant [23].

The World Health Organization has defined various diagnostic criteria allowing the classification of gliomas. Four grades of gliomas are designated according to their morphological evaluation, the proliferation index, the response to treatment and the affected patient’s survival time. Thus, the grade I includes benign tumors, the grade II includes relatively non-malignant tumors, and the grade III is assigned for the low-grade malignant tumors. As for grade IV, it is used for the most malignant tumors with a hope of life ranging from six to twelve months for the patient [62].

In medicine, imaging has become an instrumental method to help medical experts diagnose various patient diseases. In our particular application domain, glioma detection, Magnetic Resonance Imaging (MRI) has a role in identifying the tumor and determining the area in the brain for further investigation. MRI is attractive as it is non-invasive, gives high-resolution images and information unavailable with other imaging and invasive tools. In most cases, it also prevents the patient from a brain biopsy, an extremely invasive and dangerous act.

In recent years, artificial intelligence (AI) has gained significant attention from researchers in many fields, including medicine. This interest is mainly due to the perceived benefits associated with the integration of AI in medical diagnosis processes, such as higher precision, greater productivity, and faster workflow. Indeed, during the last decades, thanks to the appearance of big data, several sectors and technologies have been able to use deep learning to extract valuable information [64]. One of the different aspects of deep learning is the use of Convolutional Neural Network (CNN) for the classification of unstructured as structured data [38, 53], and especially in the analysis of medical imaging for medical diagnosis [40].

In fact, deep learning models for medical imaging have already proved to be beneficial to help medical diagnosis. For glioma segmentation, several approaches have shown their effectiveness [26, 47].

Many articles have already shown the potential of using 2D neural networks for medical diagnostics [9, 10, 40, 48]. In this work, we conducted a comprehensive analysis to determine which network is best suited to glioma pathologies. As the first diagnostic step of cancer in brain tissue, the most sensitive network will help doctors support their decision-making. However, few studies analyzed the importance of MRI sequences for choosing an effective deep learning model. Furthermore, the number of deep learning models available increases every day. It is not trivial to determine the best-performing one for a dedicated application.
One scientific contribution of this paper is to present a complete framework that investigates the performance of different imaging sequences with five effective and well-known CNN models for a particular task. We evaluate the relevance of the most usual sequences of MRI acquisition. Our study also highlights the importance of selecting and combining the best sequences for a specific task to reduce the processing and analysis time. We also study the deep features extracted by the CNN models. The deep features help interpret and understand the deep learning methods and their performance. Importantly, although our proposed framework is initially devised for glioma detection, it could be applied to other computer-aided medical tasks. Some experiments on the ADNI dataset show the adaptiveness of our framework to another task, i.e., detecting patient suffering from Alzheimer disease.

Our exhaustive experiments on three benchmark datasets demonstrate the importance of correctly choosing the sequences and the deep learning networks for the task of glioma pathologies and Alzheimer Disease. Results highlight the possible trade-off between efficiency and accuracy regarding the number of selected sequences.

Note that this paper is an extension of our preliminary study [17] with new experimental results on the Brats 2020 and ADNI datasets, showing the adaptiveness of our framework to different tasks. We also included a multi-sequence fusion study with synaptic weights and deep model correlation analysis. We also extend the analysis of the deep feature with more visualization techniques.

The rest of this article is organized as follows. In Section 2, we describe a part of the literature for computer-aided medical tasks according to the use of deep learning. Section 3 presents the material and methods, including details of our extensive framework. Experimental results on several datasets are given in Section 4, which also presents a report of the interpretability of the deep neural networks. Section 5 explains the finding of our study. In Section 6, we conclude and present perspectives of this study.

2 State of the art

For a typical examination of magnetic resonance images (MRI), doctors have at their disposal many image sequences containing different information. Navigating through the MRI data for diagnosis can take a considerable amount of time. The interpretation of MPI also relies on their experience, by which distinct neurologists may interpret the presence of glioma differently. In contrast, a deep learning approach is relatively superior for its reproducibility and stability of diagnostic and prognostic performance. Furthermore, the deep learning approach enables automated interpretation of MRI for glioma detection, working as a diagnostic assistant of pathology practitioners.

There are many types of Machine Learning (ML) applications, especially in the area of medical imaging. These applications are often classified according to the use of input data during the training of the model. At present, most common ML models employ labeled data, annotated by domain experts. From a set of input-output examples, the entire models are trained to perform specific data processing tasks.

Before the rise of deep learning, the Support Vector Machine (SVM) is a common method for MRI analysis. This method could detect and extract the main characteristics of brain tumors [2]. K-Nearest Neighbor (KNN) can also segment the brain, recognized as abnormal and presenting non-healthy tissues [31]. A less widely used method, those of the Wavelet-Entropy and Naive Bayes classifiers, could be used for feature extraction and detection of cranial pathology. The performance of this method has given rise to interesting
The current success of deep learning, a well-known Machine Learning (ML), is now attracting a lot of attention for tasks such as classification, detection, or even segmentation for such specific images. The principle of deep learning is to create an extensive network of neurons that includes many layers, with nodes within these layers. Although this method provides very high performance compared with other methods for specific tasks, this latter requires more data and more computation to train to have good results due to its sophistication. Recent advances in silicon technology, like GPU, have made it possible to use large databases for learning the deep learning model.

There are many fields of application in deep learning on MRI. Examples of possible applications are image acquisition and reconstruction (e.g., denoising and artifact detection) [6, 8, 57, 63], and improving the image resolution [3]. Among those, the applications that are of interest for us are the classification [39], segmentation [35, 42], and diagnostic assistance and prediction [41] from MRI.

For the application of glioma detection, we find two main types of studies in the literature. Those are i) the detection of glioma [11, 28, 29] and ii) the classification of the glioma by its grade [16, 22, 54]. To search for glioma in healthy brain tissue, Kalaiselvi used either the WBA (Whole Brain Atlas) database for reference [28] or a composite database between BraTS and personal clinical data for reference [29]. In the former work, the authors performed classification by wavelet methods, which concisely allowed the extraction of image features into different frequency components at different scales. In the latter, they recently developed six variants of self-customized CNN models with a five-hidden-layer architecture by varying the combinations of hyperparameter settings, such as dropout, stopping criteria, and batch normalization. Meanwhile, Chen et al. [11] attempted to extract local features slice by slice using the Histogram of Oriented Gradients (HOG) algorithm. These features representing the edge and texture of images are used by SVM with a two-phase classification for glioma detection and grading. Citak-Er et al. [16] used a personal database comprising anatomical data as well as spectroscopic data. Features were metabolic ratios obtained by MRI spectroscopy and submitted to an SVM classifier. Ge et al. [22] used a multistream CNN approach on BraTS dataset. Three CNNs were set up, each with input specific to an MRI sequence. The deep features obtained were then aggregated by a concatenation layer and then output by a fully connected layer. Shahzadi et al. [54] used a CNN VGG-16 with input from BraTS slices. The deep features extracted from the neural network were then sent to a Long Short Term Memory recurrent neural network (LSTM) to classify the glioma grade. LSTM models time dependencies and deals with the vanishing gradient problem.

In literature, it is reported that MRI sequences in deep learning are of importance for deep learning. Nevertheless, little research has shown the significance of each sequence. This work sought to fill this gap in comprehensively studying the MRI sequences for classifying two grades of glioma, i.e., healthy brain tissue and brain tissue with tumor. For example, Feng et al. [21] conducted the study of missing MRI sequences and also developed the self-adaptive neural network to deal with the missing for the lesion segmentation of multiple sclerosis. Their research eliminated the need for important MRI sequences that could not be acquired during the examination. Similarly, our study focuses on the importance of each MRI sequence with convolutional neural networks pre-trained and widely distributed with transfer learning.

The selected networks, VGG19, InceptionV3, Resnet50, DenseNet and EfficientNet-B6, participating in the ILSVRC (ImageNet Large Scale Visual Recognition Challenge) competition. The version of DenseNet used in this work is DenseNet169. We use the models and
weights derived from the Keras library, which are continuously updated. In the following, we provide the details about the selected models used in our study:

- **VGG19**: In 2014, the Visual Geometry Group team from the University of Oxford proposed a model of convolutional neural networks called VGG16 [55] and won the first and second places in the ILSVRC competition respectively in the localization and classification of the images proposed by ImageNet [51]. In this competition, they obtained an error rate of 7.4% and an accuracy of 92.7%. A new model has been proposed by the same team called VGG19. The particularity of this model is the use of extremely reduced convolutional filter layers (3x3), which allows a significant efficiency increase compared to the previous one.

- **InceptionV3**: Inception-V3 is a CNN model developed by Google [59]. The first version of the model was called GoogLeNet. He demonstrated, on the ImageNet dataset, about 78.1% of accuracy in an ILSVRC competition. The particularity of this model compared to the previous ones is that it consists of asymmetric base components including layers of different convolutions, medium pooling layers, and max pooling layers. At the end of this particular architecture, the more collected data are concatenated. Thus, each information is more collected in different ways, minimizing the risks of missing important elements.

- **ResNet50**: ResNet is a CNN model developed by Microsoft, winner of the ILSVRC challenge in 2015 [24]. The model showed an error rate of more 3.57% in the challenge of that year. The particularity of this model is that it has residual blocks called Skip Connections. These blocks are placed in order to avoid certain layers of the network, directly feeding further layers. Some information is captured by the initial layers and may be required for reconstruction and classification. Without these links between blocks, this information would be lost.

- **DenseNet**: DenseNet is a CNN model developed by Facebook AI Research [25]. The particularity of this network is that it improves the flow of information between different layers from specific patterns called Dense connectivity. Connections are introduced directly from any layer to all subsequent layers. Like the residual blocks of the ResNet network, more such connections make it possible to minimize the loss of information during training.

- **EfficientNet-B6**: EfficientNet is a recent variant of CNN models that proposes a new scaling method for identifying suitable network depth, width, and resolution [61]. The EfficientNet aims to uniformly scale all dimensions of depth/width/resolution of CNN for better performance. EfficientNet-B6 is one of the best performing EfficientNet models that transfers well and achieves state-of-the-art accuracy with fewer parameters.

These pre-trained neural networks are widely used in many fields of application. They have strong synaptic weights from the acquired knowledge. Apart from being used for the prediction task in the ILSVRC dataset, such already-built models can be employed to improve baseline performance, reduce model-development time, and optimally utilize existing knowledge to solve different target tasks. This method is generally called transfer learning, by which a model trained to solve one problem can be applied to solve another but related problem. For example, knowledge gained while learning to recognize bicycles could be used when trying to recognize motorbikes. In the medical field, ResNet50 was employed to leverage the formation of brain tumors to detect abnormal methylation in the brain tissue [36]. The dense connectivity of Densenet could be more fine-tuned to improve MRI image resolution [13]. For the detection of Coronavirus Disease, the ResNet50 and...
inceptionV3 networks were applied by further training and testing them with images from X-ray chest radiography [45]. For a more complex medical imaging technology, 3D volumetric imaging allows a surgeon to see an accurate picture of human anatomy, analysing the issue at hand before picking up a scalpel. Accordingly, 3D neural network models are also designed for learning dense volumetric segmentation from sparse annotation [15]. At present, they are also widely used in the ML field of medical images for various applications, such as segmentation of the pancreas [46], cell counting [20], and, which is our focus here, the segmentation of brain tumor [12].

More similar to our study, the study of Hwan-ho Cho et al. [14] used BraTS 2015 to perform multi-level grade glioma classification. They also applied under-sampling to obtain a balanced dataset and searched for Histogram-based features, Shape descriptors, GLCM-based features via a min squared error (MSE) algorithm of LASSO. Saed Khawaldeh et al. [33] used convolutional neural networks on a personal glioma database for the classification between Healthy, LGG, and HGG, labeled according to the recent new criteria WHO (2016). Finally, Muhammad A. Khan et al. [32] employed the marker-based watershed algorithm and multilevel priority feature selection for detecting and classifying brain tumors on three different databases, including BraTS 2013, Harvard, and private.

3 Material and methods

In this section, the pipeline of our study for glioma classification is described. Figure 1 illustrates the analysis workflow, starting from the MRI database of brains to detect glioma. 3D high-resolution images of patient heads were undergone skull stripping and then extracted for the sequences of 2D slices and 3D cubes. Next, we fetched these different images to a pre-processing step of image normalization. We subsequently applied a data augmentation technique before feeding them to the 2D and 3D pre-trained neural networks according to their dimensions. Finally, the transfer learning approach was employed to incrementally fit the pre-trained convolutional neural networks for the glioma classification task. After this step, we made an exhaustive analysis of the performances on the combinations of modalities and networks. At the same time, deep features were extracted and analyzed to make an interpretability analysis. This investigation of feature maps provides a better understanding of the features used by networks for their learning.

3.1 Data

In medicine and medical research, MRI has become a vital diagnostic method for detecting glioma. Through this optical imaging modality, several modalities for sequence acquisition are used to highlight different tissues of the tumors. Examples of those notable tumor tissues are edema, necrotic core, and active edge. For the detection of edema, the T2 FLAIR (Fluid-attenuated inversion recovery) sequence can be used. Likewise, for the detection of the tumor core, often necrotic tissue, the T1ce sequence (a T1 sequence coupled with injections of gadolinium as a contrast agent) can be applied.

We selected the BraTS challenge 2018 and 2020 databases [4, 5, 43] that contains different sequences for our study. The BraTS database (Multimodal Brain Tumor Segmentation) is a database produced by CBICA (Center for Biomedical Image Computing and Analysis) in partnership with MICCAI (Medical Image Computing and Computer Assisted Intervention Society). Brats 2018 database contains MRI brain data, collected from over 300 patients.
Fig. 1 Pipeline of the current study
scanned in a transverse plane. The data is in the format NIFTI file (.nii.gz), for four different sequences, listed as follows:

- **T1 sequence**: Longitudinal relaxation
- **T1ce (T1-Contrast Enhanced) sequence**: Longitudinal relaxation with addition of contrast medium (gadolinium)
- **T2 sequence**: Transverse relaxation
- **T2 FLAIR (Fluid Attenuated Inversion Recovery ) sequence**: Transverse relaxation with inversion recovery (attenuation of fluids)

Figure 2 illustrates the four types of sequences scanned from a patient with a high grade of glioma.

As obtained from 19 different institutions, this MRI database contains different grades of glioma. The segmentation of tumors and their labeling were done manually beforehand by following the same protocol of approved and experienced neuroradiologists. According to the description of the BraTS 2018 database, these data represent 210 patients with high-grade gliomas versus 75 patients with low-grade gliomas. However, according to local neuroradiologists and according to WHO criteria, as shown by Dequidt et al. [18, 19], 254 patients would be high-grade patients and 31 low-grade patients. Thus, the final ratio of LGG to HGG would be 0.109, as reported by the authors, making the study very complicated to balance with this type of database. So, for our study and the research of glioma, we use deep neural networks to classify and predict the presence or absence of glioma. Thus, two outputs are present: only healthy brain tissue and brain tissue with the presence of tumors. These data were divided into training, validation, and testing sets. They then underwent pre-processing steps for normalization, preparation, and data augmentation before ultimately being used by CNNs.

To carry out the study, we also selected the BraTS 2020 dataset. This database differs from the previous ones on several points. Only images and annotations from BraTS’12-'13 were kept. The other data were removed because they contained a mixture of pre and post-operative analyses. Neuroradiologists radiologically reassessed all of the original TCIA glioma collections. Accordingly, all the pre-operative scanners were annotated by experts and included in the BraTS dataset 2020. In total, our study experimented on the data from 369 patients, consisting of 259 with high-grade glioma and 110 with low-grade glioma. Moreover, 84 MRI sequences were added to BraTS 2020 in comparison with BraTS 2018, i.e., +29.6% of new data.

Finally, to generalise the study, we carried out the workflow on another small dataset: ADNI. We selected two subgroups of twenty patients, consisting of ten healthy and ten...
severely affected by Alzheimer’s disease. Three different MRI sequences (i.e., MPRAGE-TSE, Double-TSE, and T2-Proton Density) common to the subjects were selected for each patient.

To comply ADNI dataset copyrights, we include the following sentences: “Data used in the preparation of this article were obtained from the Alzheimer’s Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). The ADNI was launched in 2003 as a public-private partnership, led by Principal Investigator, Michael W. Weiner, MD. The primary goal of ADNI has been to test whether serial magnetic resonance imaging (MRI), positron emission tomography (PET), other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of mild cognitive impairment (MCI) and early Alzheimer’s disease (AD).”

3.2 Pre-processing

Before training the neural networks on the selected database, the NIFTI files (MRI) are converted into image data to cope with the CNN input. Generally, the Keras library accepts many image formats (e.g., PNG, JPG, BMP, PPM or TIFF) for training neural networks. For our study, we use the PNG format because of its lossless compression\(^1\). Besides, experimental images must contain three channels corresponding to the base of the RGB channels. This requirement is due to the fact that the studied networks can only accept images in the RGB format while some with a minimum pixel size are required.

This idea of pre-processing MRI data was already carried out and validated by Ahmad et al. [1] for muscle segmentation by deep transfer learning. Furthermore, before normalizing the data, the intensity scaling (IS) of all brains was applied for only healthy regions of brains. The IS method is described in Xiaofei Sun et al. [58].

This paper also compares different ways to normalize histograms of images from MRI, including the input image, histogram matching, and histogram normalization. Acquiring images from multiple machines, different protocols, and distinct manipulators results in the great heterogeneity of the histograms of acquired images. To overcome this bias and recover a normalization to better help CNNs to learn, many histogram normalization approaches [7, 27] were tested and the latter is to date one of the most efficient. Following this, a min-max normalization was applied for the intensity normalization of voxels.

The pre-processing of data is done in several steps:

1. First, the IS is applied with the following the (1):

\[
N(x, y, z) = \begin{cases} 
\mu_s + (f(x, y, z) - \mu_i) \frac{LIR - \mu_s}{S_{1i} - \mu_i}, & \text{if } m'_1 \leq f(x, y, z) \leq \mu_i \\
\mu_s + (f(x, y, z) - \mu_i) \frac{HIR - \mu_s}{S_{2i} - \mu_i}, & \text{if } \mu_i \leq f(x, y, z) \leq m'_2 
\end{cases}
\]  

where LIR (low intensity region) is the lowest decile of the histogram of the standard image and HIR (high intensity region) the highest decile and \(S_{1i}\) is the lowest decile of the histogram of the input image and \(S_{2i}\) the highest decile.

2. Then, each of these sections is normalized according to the min-max normalization equation presented in (2).

\[
X_{scaled} = \frac{X - X_{min}}{X_{max} - X_{min}} \tag{2}
\]

\(^1\)Although the TIFF format is also suitable, its files tend to be relatively quite large.
3. For each patient, we get their transverse sections that the neural networks will use to train. According to the segmentation from the experts, each transverse section is classified and labeled whether it has a glioma or not. Slices with only background values (black) are excluded; they represent the outer limits of the cranial box where no data is present. These slices are converted to PNG format to match the input format of 2D NNs. Converting images to PNG thus reduces the depth of the voxels from 16 bits to 8 bits. Information can therefore be lost. One of the objectives of this study is to find out whether the information contained in such reduced dimensions may be sufficient for CNNs learned heavily from standard images. Thus training on such images can significantly reduce computer resources and processing time by using the transfer learning technique. This procedure is confronted with the study of a 3D network, remaining in 16-bit format. Due to the large dataset (50,812 slices), we used a three-way hold-out method to split data into three different subsets (i.e., training, validation, and test sets). The data was earlier balanced and divided such that 70% are used for training, 15% for validation, and 15% for testing. Note that this splitting was not performed randomly but was performed on the patient and not on the slice. This way, we could ensure that the training, validation and testing sets are not biased.

3.3 Data augmentation

In general, to get a good performance out of machine learning models, we need a good number of examples proportional to the number of parameters of the models; and the complexity of the tasks the models have to perform. Due to the scarcity of our study data, it is necessary to apply an image data augmentation technique to enlarge our dataset. Also, using this technique does not only unravel the limited data but also helps deep learning algorithms avoid the problem of overfitting. Although CNN can be invariant (i.e., images can be placed in different orientations, translation, viewpoint, size or illumination), medical datasets still contain samples in a limited set of conditions. In contrast, MRI in real-world situations may exist in a variety of conditions, such as different orientation, pose, scale, brightness, etc.

Before feeding training data to neural networks, we performed several transformations on the original MRI data. All the necessary transformations we applied in this work are as follows:

- rotations up to 40 degrees;
- longitudinal deformations up to a ratio of 0.2;
- axial deformations up to a ratio of 0.2;
- shear effects up to a ratio of 0.2;
- zoom up to a ratio of 0.2;
- horizontal and/or vertical flips.

3.4 Implementation of models

For each of the models, their synaptic weights were loaded according to their performance on imageNet for the study [51]. These 2D networks required tremendous computing power for training and colossal time to achieve their performance results for the ILSVRC competition. These different 2D neural networks have between 17 and 31 million trainable parameters (43 million for EfficientNet-B6). Training each parameter on a small MRI database and hoping to adjust synaptic weights for these millions of parameters is illusory.
The goal of transfer learning is to preserve these previously learned synaptic weights for these neural networks on millions of images from the imageNet challenge and, as a result, freeze the layers and make the parameters non-trainable. The main idea behind this approach is that the characteristics of a medical image, whether it comes from an MRI scan or other modalities, have the same characteristics as a “standard” image. Thus, the shapes, the differences in intensity between neighboring pixels, the shades of colors or gray, etc., which allow the classification of the ILSVRC challenge, are found within the MRI images.

Tajbakhsh et al. [60] reported that transfer learning showed great potential for specific tasks for databases with very few annotations. Only the bottlenecks, which represented the output of these models, were changed to be task-specific. Accordingly, we removed the last layer in each of the studied models and added a flatten layer to standardize the data into a single vector. This layer was subsequently connected to a small network of neurons, forming fully connected neurons. This removal possibly introduced the dropouts of 0.5, meaning that only 50% of neurons pass the learned information during the training phase. By doing this, we regularized the deep neural networks by not allowing the neurons to pass the information to the next layer. This method thus reduces the possibility of overfitting and improves generalization error. Finally, the output of these neural networks was linked to a final layer, made up of as many neurons as there were classes during the study. Here, two neurons were present. Each represented a classification output, which means, a section of the brain representing only healthy tissue, or a section of the brain with the presence of a glioma. It was only these last few layers trained on the annotated images of the MRI database. The total number of parameters that we could train are presented in Table 1, and Adam 10-5 was the optimizer.

The dropout layers allow the training to be stopped for a certain percentage of neurons, preventing the readjustment of their synaptic weights at each epoch. The main purpose of dropout layers is to avoid overfitting and enhance the generalization ability of the networks. Without these dropout layers, the InceptionV3, ResNet50, and DenseNet models were close to 100% accuracy on the training set but a completely random performance on validation or test sets, which is an apparent overfitting problem.

The fully connected layers regroup the extracted characteristics from convolutional blocks to propose a classification. All the neurons in those layers are interconnected and have their own weight. The use of such layer is delicate because the number of parameters increases exponentially with respect to the number of neurons present in the layers. Thus, the computation time increases accordingly. The general scheme of the implementation of the models is illustrated in Fig. 3.

**Table 1** Number of parameters for each model and specific features

| Model     | 2 outputs | total         | not trainable | trainable | specific features                  |
|-----------|-----------|---------------|---------------|-----------|------------------------------------|
| VGG19     | 21,630,210| 20,024,384    | 1,605,826     |           | Reduced convolutional filter       |
| InceptionV3| 26,521,570| 21,802,784    | 4,718,786     |           | Asymmetric components              |
| ResNet50  | 31,976,514| 23,587,712    | 8,388,802     |           | Skip Connections                   |
| DenseNet  | 17,861,378| 12,642,880    | 5,218,498     |           | Dense connectivity                 |
| EfficientNet-B6| 50,397,514| 40,960,136    | 9,437,378     |           | Uniformly scale dimensions         |
4 Results

In this section, we start by studying the existing correlation between MRI sequences and CNNs. Next, we conduct a comprehensive experiment to see which model or sequence is more pertinent to glioma prediction. We thus mix and train all investigated models and methods for a thorough analysis of the prediction performance of their combination. Afterward, we compare the 2D neural networks with the UNET3D network, widely used today in augmentation from medical imagery. Finally, we proceed with the extraction of deep features and analyze them to improve the interpretability of the data.

4.1 Performance and correlation phenomenon

Our study seeks to help the medical expert in the first step of the analysis, i.e., detecting tissue present a glioma by classification using a convolutional neural network (CNN). Our study aims to evaluate the effectiveness of different CNNs for different MRI sequences. Therefore before starting the study, it is necessary to verify the existing correlation between sequences and networks.

In this first set of experiments, we tested four convolutional neural networks, i.e. VGG19, InceptionV3, ResNet50 and DenseNet, on each MRI sequence. Their predictions were then exploited to serve as inputs to a Dense-2 layer, as shown in Fig. 4 (a). The synaptic weights of each branch were also analyzed. Figure 4 (b) depicts the weight analysis, for example, performed on T1 MRI sequence.

Once the best combination of sequences and neural networks had been found and recorded, we repeated the same analysis. Indeed, particular sequences could provide similar information, i.e., their performance could be hidden by another sequence performance when these two models were used together. Therefore, this step was repeated several times in order to more ensure that the weights of the perceptrons remained in the same order.

The analysis showed that the order of the perceptron weights did not change. This result implied that a sequence or a neural network systematically brought few but different information. The maximum accuracy was 0.8638 when all networks and all sequences were combined.

Then, the best couple of weighted network sequences (as highlighted in bold in Table 2) were removed one by one until only eleven couples of weighted network sequences were left after removing ResNet50-T2. As we can see in Table 2, the accuracy gradually decreases one after the other.
We then compared the performance of all selected CNNs for a fixed sequence. Thus, we were able to determine the most significant weights for all sequences. This experiment is very useful to determine the CNN mostly suited for our study. An example of the T1 sequence is shown in Fig. 4 (b).

As we can see in Table 3, the T2 FLAIR is the best sequence (highlighted in bold), which has better accuracy than the T2 sequence, followed by T1ce and T1 sequences. We conjecture that the importance of a network is determined by the absolute difference between Healthy and Gliomas weights for one CNN. The most important weights are of DenseNet and Resnet50 models for the T2 FLAIR and T2 sequences. VGG19 is the model with the lowest weights and would therefore be the least informative model for this kind of study when combined with others.

Next, we attempted to determine which an MRI sequence, or a combination of several sequences, was more effective than others for the first stage of diagnosis, i.e. the detection of a glioma. By doing so, we trained each of the four studied models on each possible combination of sequences. The results were then sorted by the model, which obtained the best accuracy.

![Diagram](image)

**Fig. 4** Global weights analysis (a) and weights analysis by MRI sequences (b)

| Model and sequence deleted | Accuracy |
|----------------------------|----------|
| all sequences and models   | **0.8638** |
| without ResNet50-T2FLAIR   | 0.8560   |
| without DenseNet-T2FLAIR    | 0.8464   |
| without InceptionV3-T2FLAIR | 0.8358   |
| without VGG19-T2           | 0.8335   |
| without ResNet50-T2        | 0.8253   |

*Table 2* Accuracy after the sequential deletion of model and sequence from the best couple
The highest accuracy is obtained with ResNet50 (85%), closely followed by DenseNet and InceptionV3 models (84%). Note that VGG19 model obtains significantly lower accuracy scores than the first three (around 79%). In addition, we can notice that a significant drop, around 2 to 3%, in terms of accuracy occurs when the T2 FLAIR sequence is removed from the sequences. From these results, we can determine that the best model is ResNet50 for our particular application and that the T2-FLAIR sequence is of absolute importance for this specific medical task.

4.2 Mixing all

This section describes the results of our comprehensive analysis for the performances on the combinations between MRI sequences and neural network models on Brats 2018 dataset. This analysis would help determine the predominant couple of the sequence and model for classifying the presence of glioma in a patient brain.

All possible combinations were tested and examined, and their predictions were then used as inputs to a Dense-2 layer. In other words, there were a total of 16 different possible inputs, derived from four different models and four different sequences, for investigation.

We subsequently tested the best combination of 2, 3, and 4 out of the 16 (i.e., 120, 560, and 1820 different combinations, respectively.) According to their performance, these combinations were sorted in descending order, ranging from 0.86 to 0.76. Figure 5 demonstrates the results of the combinations of the four neural networks (i.e., Resnet50, DenseNet, InceptionV3, and, VGG19), grouped by the bin of 10, 20, and 35 combinations for visual purpose respectively for Sub-Figures (a), (b), and (c).

A similar test was carried out for the four CNN models. Observed results were similar to the previously obtained results in Section 4.1.

We noticed, in each case, there was a clear predominance of the T2 FLAIR sequence (in orange), i.e., the greatest frequency to be found in the combinations obtaining the highest accuracy. When this T2 FLAIR sequence was not considered any more, the T2 sequence (green) compensated for the loss of information that the T2 FLAIR sequence provided. Finally, when the T2 FLAIR and T2 were no longer active, T1ce took over, then finally T1. As a result, we can determined a clear ranking of the importance of the different sequences for this study.

A similar experiment was also performed on the BraTS 2020 database to confirm our findings. Additionally for part of the study, the neural networks EfficientNet-B6 was implemented. Figure 6 presents perfectly similar results with the previous database, here we tested...
Fig. 5 Frequency usage of each sequence according to accuracy of the combination (a) 2 sequences out of 16, (b) 3 out of 16 and (c) 4 out of 16. All combination were sort decreasingly by accuracy. In y axis, the number of occurrence of these modalities found and in x axis the number of the combination for the four different CNNs: Resnet50, DenseNet, InceptionV3, and VGG19 and the four different sequences : T1, T2, T2FLAIR, and T1ce
Fig. 6 Frequency usage of each sequence according to accuracy of the combination 3 out of 20. All combinations were sort decreasingly by accuracy. In y axis, the number of occurrence of these modalities found and in x axis the number of the combination for the five different CNNs: Resnet50, DenseNet, InceptionV3, EfficientNet-B6, and VGG19 and the four different sequences: T1, T2, T2FLAIR, and T1ce for the database BraTS 2020

the best combination 3 out of 20 (i.e. 1140 different combinations). This figure demonstrates the results of the combinations of the five neural networks (i.e., Resnet50, DenseNet, InceptionV3, EfficientNet-B6 and VGG19), grouped by the bin of 20 combinations for visual purpose. For this analysis, like the previous networks, EfficientNet does not make a significant change in the order of performance. Here again, the type of sequences remains decisive in the accuracy. For this glioma detection study, there was a preference in decreasing order for the T2 FLAIR, T2, T1ce, and finally T1 sequence. Note that, in general, the obtained accuracy for this database was higher by one or two percent for this database. This increment is because, in the BraTS 2020 database, experts provided a cleaner dataset. Even if both datasets have differences (30% of different patients), they share common MRI resulting in similar conclusions in all our experiments.

4.3 RGB re-composition from MRI sequences

As we demonstrated previously, some MRI sequences seemed more effective than others for determining the presence of a brain tumor via the use of transfer learning from 2D pre-trained convolutional neural networks. Among these sequences, we found in preferential order, i.e., T2-FLAIR, T2, T1ce, and finally T1 sequence. Therefore, we accordingly tried to combine these

---

2Some brains of patients who underwent postoperative operations were removed. This removal might help the training in general.
sequences in a re-composition of an RGB image to use these trichromatic images as a learning set. Figure 7 depicts the combination of sequences within the same image. T2 arbitrarily takes the red channel, T2-FLAIR takes the green channel, and T1ce takes the blue channel.

Figure 8 illustrates the evolution of the accuracy and the loss function for the best two selected CNN models, i.e., ResNet50 and DenseNet.

The accuracies for the DenseNet, InceptionV3, ResNet50 and VGG19 networks were respectively at 0.855, 0.865, 0.835, 0.820 for the validation data. The fact that the training data had lower values than those of validation during the first iterations was explained by the data augmentation. Indeed, as stated previously, only the training data was augmented. By doing this, we prevented the models from over-fitting, and thus allowed the models to focus on the characteristics for a better classification. This over-fitting was also observed on the loss data. Delaying this phenomenon or improving the whole learning process could be envisaged with the use of more data for training.

The loss data was 0.31, 0.43, 0.40 and 0.44 respectively for the DenseNet, InceptionV3, ResNet50 and VGG19 models. These different data showed better performance for the DenseNet and ResNet50 models. The latter were also those with the least loss, which confirmed the previous observation.

4.4 Comparison with UNET3D and literature survey

As previously specified in the state-of-art section (Section 2), UNET-3D model [49] has become prevalent and popular when the goal is to achieve a specific task in 3D medical imaging. This kind of model is a network that specializes in medical task segmentation; hence, we can easily adjust it for a detection and classification task. In this section, we, therefore, compare the UNET-3D with all of the 2D CNNs selected for this study.

To perform the comparison, we cut each brain in the database into cubes of 32x32x32 voxels. Any cubes containing more than 20% of black voxels were excluded to improve the quality of the training set due to their lack of information. Whereas any cubes had more than 15% of voxels belonging to the glioma, we then labeled them as diseased brains. In contrast, if they had less than 1% of voxels presenting a glioma, the cubes were considered healthy.

The UNET-3D model was trained using all possible combinations of sequences. The model was then truncated to keep only the convolution part, added to a dense (fully connected) network of 64 perceptrons, followed by a dense layer for our detection task. The pre-learned parameters of the basic model were frozen, and the fully connected layers were trained with the previously described datasets.
As we can see in Fig. 9, whatever the sequence is, the UNET-3D model is more efficient and has better accuracy than 2D networks, with the maximum accuracy of 0.84 for the combination of the three sequences, i.e., T1, T2, and T2-FLAIR. Note that there are performance changes for the same combination of sequences, of which the network can be trapped in local minima. On average, the performance for each combination is relatively
During the prediction of this model, the cubes with gliomas between 1% and 14% were added to the experiment, unlike training. If these latter cubes had not been added, the performance would have increased to 93% for the best combination. Likewise, if the 2D slices tangential to the glioma had been removed during the training and the prediction, the accuracy of the 2D neural networks would have gone up equivalently. From the results obtained, we can hypothesize that the depth reduction of the image from 16 bits to 8 bits does not influence, or not significantly enough, the performance of the networks used. It should be noted that this concerns overtraining networks coming from ImageNet challenge, on millions of images.

We produced confusion matrices for the predictions on the test set. Detecting the presence of glioma is of initial importance. Indeed, for a glioma computer-aided diagnosis, it is better to have false positives (330) than false negatives (38). As shown in Fig. 10, we observe, for example, only 38 poorly predicted cubes for the values presenting a glioma for the trained model having the highest accuracy for the combination of sequences. As each patient’s brain was cut between 100 and 150 cubes, the results with only 38 cubes misclassified (on different patients) would not change the diagnostic of the patients; no glioma would be left undetected. Note that those miss-classified cubes were mainly located on the edge of the glioma, with a tiny percentage of voxels being part of the glioma.

![Confusion matrix for UNET3D for the combination of T2 FLAIR, T1 and T2 sequences. HGG is the label for the presence of a glioma, NG is for healthy brain](image-url)

**Fig. 10** Confusion matrix for UNET3D for the combination of T2 FLAIR, T1 and T2 sequences. HGG is the label for the presence of a glioma, NG is for healthy brain.
To complete our study, we conducted a performance comparison extracted from recent literature approaches detailed in Section 2. Still, we remind the reader that our study objectives are to show our framework highlights the efficiency of pre-trained CNN models and understand the importance of MRI sequences for a specific disease.

Table 4 reports the accuracy performance of recent literature dedicated approaches on MRI sequences with respect to the classification type: “glioma vs healthy” (top of the table) and “glioma grade” and the dataset used. As expected, recent and glioma dedicated models show better performance than our framework, which performance is fairly close. However, most approaches shown in Table 4 have their own classification goal and selection of data from the Brats dataset or their own dataset. For example, Kalaiselvi et al. [29] used the Brats2013 dataset only as training set while using the WBA dataset in both training and test sets. It is therefore hard to compare accurately all performances across different experimental settings.

4.5 Study of the workflow on an additional database: ADNI

In this part, we reused the workflow with the 2D convolutional neural networks (i.e., Resnet50, Densenet, inceptionV3, VGG19, EfficientNet-B6) previously used with another database, ADNI. ADNI has many patients with or without Alzheimer's disease. This database has the particularity of offering many different MRI sequences. For this analysis, we selected 20 patients, consisting of ten healthy and ten severely affected by Alzheimer's disease. Three different MRI sequences (i.e., MPRAGE-TSE, Double-TSE, and T2-Proton Density) common to the subjects were selected for each patient. Due to the small quantity of the selected subset, the leave-one-out method was chosen for the distribution of the dataset between train and test sets. For each patient, the same pre-processing steps were applied. Distal cerebral slices were excluded due to their lack of data. Only sections showing an encephalic part were preserved. Whatever the MRI model-sequence combination was, the performance was always above 0.9 on average. The only exception was for InceptionV3 with the MPRAGE modality, which could not validate its training for the chosen parameters.

The analysis here shows that the choice of modality depends entirely on the selected network. Indeed, For the 100 best combinations as shown in Fig. 11, the MRI sequences are Double-TSE and Proton-Density-T2 in an equivalent manner, in combination with the DenseNet and EfficientB6 models. Following this, the best combination is with MPRAGE,

| Works                  | Model          | Classification       | dataset          | Acc.   |
|------------------------|----------------|----------------------|------------------|--------|
| Ours                   | DenseNet       | healthy vs glioma    | BraTS            | 85.5%  |
| Ours                   | ResNet50       | healthy vs glioma    | BraTS            | 86.5%  |
| Ours                   | InceptionV3    | healthy vs glioma    | BraTS            | 83.5%  |
| Ours                   | VGG-19         | healthy vs glioma    | BraTS            | 82.0%  |
| Kalaiselvi et al. [28] | Wavelet Method | healthy vs glioma    | WBA dataset      | 89.6%  |
| Kalaiselvi et al. [29] | CNN            | healthy vs glioma    | BraTS + WBA dataset | 88.9% |
| Chen et al. [11]       | HOG + SVM      | healthy vs glioma    | BraTS            | 76.8%  |
| Citak-Er et al. [16]   | spectro + SVM  | glioma grading       | own dataset      | 93%    |
| Ge et al. [22]         | Multistream CNN| glioma grading       | BraTS            | 90.87% |
| Shahzadi et al. [54]   | VGG-16 + LSTM  | glioma grading       | BraTS            | 84%    |
Fig. 11 Frequency usage of each sequence according to accuracy of the combination (a) 3 sequences out of 20, (b) Frequency usage of each neural network according to accuracy of the combination (a) 3 sequences out of 20. In y axis, the number of occurrence of these modalities found and in x axis the number of the combination
combined with the ResNet neural networks. The hypothesis that we can advance here is that each model is more suitable for training according to the sequences, and that by this fact, each model better analyzes certain features specific to the different sequences.

We can thus think that for the best analysis, with the best accuracy, it is mandatory to choose the combination of models and MRI sequences according to the pathology.

4.6 Deep feature extraction and interpretation analysis

In this section, we propose to analyze the deep features extracted on the end of each CNN used in this study for each MRI sequence. Trying to understand and interpret CNN has become an active research field. Researchers have attempted to open the black box to understand how or why the prediction has been decided. Among the numerous articles of this explainable artificial intelligence field, Lapuschkin et al. [37] and Samek et al. [52] try to demystify the CNN.

Inspired by their work, we decided to put our efforts on Layer-Wise Relevance Propagation (LRP) [44], Deep Taylor decomposition [30], and the Guided Backpropagation [56] methods to highlight the interesting features obtained by each sequence and network. The general aim is to deduce biological rules within different descriptors.

We first had a look at the visual difference between feature maps of the three selected methods for MRI axial slices, as predicted by the four CNN models. Figure 12 depicts examples of the visual map results on different patient slices. Images in the first row are from the Epsilon-LRP method tested on trichromatic images (T2-FLAIR, T2, and T1ce sequences for red, green, and blue, respectively). In the second row, the Deep Taylor method was applied to the T2-FLAIR sequence. The last row shows the examples of the Guided Backpropagation method tested on the T1 sequence.

In the following, we give general observations for the three visualization methodologies. For the Epsilon-LRP method (as we can observe in the first row of Fig. 12), the parenchyma and the cerebral periphery are mainly highlighted, similar to a homogeneous noise. On the contrary, we find the main structures, such as the ventricles, in heterogeneous whiter spaces. The edges of the contours of the main structures seem more marked for DenseNet and ResNet50 CNN. In the second row of Fig. 12, we observe the results of the Deep Taylor method. In this specific example, each neural network could detect the outer bounds of the lesion materialized by the T2 FLAIR edema. We note that for this image, all networks would perform well. However, some of them (VGG19, InceptionV3) better detected the borders of cerebral ventricles. We could explain this outcome by a well-known periventricular artifact in the T2 FLAIR sequence. This artifact could mimic the edema part of the tumor and lose the network in this classification task.

With the T2 FLAIR sequence, we could easily visualize the presence of a rupture within the image symbolizing the limits of tissues, such as cerebral convolutions, for all CNN models. We could also visualize cerebral furrows, ventricles, and, in the event of the presence of a tumor, the edematic part.

Particularly note that the ResNet50 model could detect the presence of specific texture for the internal part of the glioma edema. In contrast, the other models could only detect a scission between the two tissues. This finding could help provide a better classification.

In the last row of Fig. 12, the visualization of deep feature maps for the Guided Backpropagation method is presented. Note that for the Fig. 12j, the output image (from Inception) was almost imperceptible. An adjustment of the brightness and the contrast was applied to visualize the feature map better. We noticed that this last method focused on large struc-
tures, forming a clear break in the original image. Thus, the cerebral periphery, the scission between the cerebral hemispheres, was visible. All the different studied neural networks seemed to have similar behaviors for this method with T1 sequences.

We could find a lot of similarities between the different couples of methods and CNN models. Although they all reacted similarly (i.e., they extract mostly the same information), some of them had specificities. They also brought a particular analysis for some details. For instance, the ResNet50 on the Deep Taylor method specialized in extracting the texture of the edema part. The DenseNet and ResNet50 on the Epsilon-LRP method could highlight the edges of the structures relatively more. This analysis confirmed each CNN model brought its own additional information from the different MRI sequences.

For the second part of visualization analysis, we studied the feature maps extracted for one neural network but different MRI sequences. We thus selected the ResNet50 model that performed best in our prior study (Section 4.1). Feature maps are presented in Fig. 13.

---

**Fig. 12** Deep feature maps of Epsilon-LRP on a trichromatic (T2-FLAIR, T2, T1ce) image (row 1), Deep Taylor on T2-FLAIR sequence (row 2), Guided Backpropagation on T1 sequence (row 3), using DenseNet (column 1), InceptionV3 (column 2), Resnet50 (column 3) and VGG19 (column 4)
The input images are the re-composed trichromatic images (RGB visualization) in the four different sequences, i.e., RGB, T1, T2, T1ce, and T2 FLAIR (in row order). Each column visualizes the input images and the feature map results of the Epsilon-LRP, Deep Taylor, and Guided Backpropagation methods (from left to right).
In the following, we give general observations for the three visualization methodologies. For the Epsilon-LRP method, we found similar observations to the previous figure. In addition, in the second column of Fig. 13, in the glioma region, the tumor core of the high-grade glioma was mainly observed in heterogeneous white regions.

The Deep Taylor method seemed to give the most accurate visualization. A clear homogeneous texture represented the parenchyma. Structures, such as cerebral convolutions and furrows, ventricles, and the different parts of the glioma, were eminently highlighted.

The last method, Guided Backpropagation, also gave similar results as presented in Fig. 12. This method could identify large structures, which generally are the cerebral periphery, scission between the cerebral hemispheres and ventricles, and all the different tissues of the tumor.

We could also confirm the previous conclusions: the asymmetry in the feature map given by the three methods may be used to identify healthy to non-healthy brains. The most relevant images seemed to be the trichromatic image as they gathered multiple parts of information from three MRI sequences. The resultant feature maps gave the best visual information on the brain.

Finally, we compared the feature maps of the different MRI sequences. Importantly note that to find the necrotic core, the T1ce is the most interesting sequence. The leak of contrast liquid (gadolinium) brought a hyper-signal around the necrotic core where the tumor produced a neo-vascularization. Thus, when there was a limit of the necrotic core, the tumor cells had a very strong activity. However, glioma without a necrotic core would be difficult to detect. The T2 and T2 FLAIR sequences seemed to be more involved in detecting the edema area, which corresponded to the entire volume of the glioma. The T1 sequence was likely to be more involved in detecting of the whole tumor part without actually identifying the necrotic core. Each sequence, therefore, had its importance in the detection of the different histological parts of the glioma.

5 Discussion

In this section, we summarize the findings of our study about our main application, the glioma detection. The first essential information is that within the same MRI, CNN models extract different types of information according to the nature of the sequence. This observation can indeed be extended for any imagery analysis based on MRI when applying deep learning. There is a part of information uniquely provided by each sequence but unable to be replaced by any other. Therefore, to answer the question: “Can we absolve ourselves from using one type of sequence without losing essential information?”; the answer is “No, we cannot use only one type of sequence.” Removing information would save valuable time, but removing it could significantly impact the performance of the computer-aided diagnosis. However, not all of them are effective for any dedicated application. Selecting particular modalities or sequences requires an elaborated experiment like what we did in this study.

Following our work on the four well-known CNNs with a transfer learning strategy, we are able to observe a slight preference for three of them (i.e., ResNet50, DenseNet, InceptionV3), reacting similarly. The last one, VGG19, seems to be significantly less efficient for this specific task.

A key piece of information for the study of glioma detection against healthy tissue is the importance of the T2 FLAIR sequence. For each of the three preferred neural networks,
all the combinations providing the best performance are those including the T2 FLAIR sequence.

There seems no essential information lost when we exclude the T1 sequence from trichromatic images for gliomas identification. The trichromatic images that consist of just the T1ce, T2, and T2 FLAIR sequences provide the best performance in our study. The discussions with medical experts also confirm their preference for these three MRI sequences as neurosurgeons also use them to analyze gliomas.

The visual analysis of deep features could show a histological interest of the models according to the sequences. For example, the T1ce sequence has lower performances than other sequences as it is dedicated to the lesion contrast enhancement, only identifying high-grade glioma. In comparison, the T2 FLAIR finds edema in all gliomas (low and high grades). To sum up, each sequence type allows visualizing specific histological characteristics similar to what medical experts focus on for brain diagnoses.

6 Conclusion

This article introduced a global framework for particular medical applications. We proposed an exhaustive performance study of five well-known 2D CNN and a 3D CNN, specifically used for medical images. All these models were pre-trained and utilized together with transfer learning on multimodal sequences from MRI. We were able to show the importance of mixing sequences to improve the overall performance. Depending on the medical task, some sequences are mandatory, but others are not. Nevertheless, many sequences carry their unique parts of information that should not be neglectable without a comprehensive study. Even if our study is mainly applied on the detection of glioma, we show the adaptiveness of our framework to another medical task, detecting Alzheimer Disease, with different types of sequences. We also showed the importance of selecting a well-adapted convolutional neural network model depending on the application. Finally, we proposed a visual analysis and interpretation of the multimodal features obtained from the different CNN. Therefore, medical specialists could have a better understanding of results obtained by deep learning models. Our future work will focus on the improvement of the interactivity between the visualization tool and the medical experts to enhance the learning process of our framework.

Acknowledgments  This work was supported in part by the Academic Melting Pot Program from the King Mongkut’s Institute of Technology Ladkrabang (KMITL) for the fiscal year 2019.

Data collection and sharing for this project was funded by the Alzheimer’s Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904) and DOD ADNI (Department of Defense award number W81XWH-12-2-0012). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: AbbVie; Alzheimer’s Association; Alzheimer’s Drug Discovery Foundation; Araclon Biotech; BioClinica, Inc.; Biogen; Bristol-Myers Squibb Company; CereSpir, Inc.; Cogstate; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; EuroImmun; F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; Fujirebio; GE Healthcare; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Lumosity; Lundbeck; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; NeuroRx Research; Neurotrack Technologies; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Takeda Pharmaceutical Company; and Transition Therapeutics. The Canadian Institutes of Health Research is providing funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health (www.fnih.org). The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer’s Therapeutic Research Institute at the University of Southern California. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of Southern California.
References

1. Ahmad E, Goyal M, McPhee JS, Degens H, Yap MH (2018) Semantic segmentation of human thigh quadriceps muscle in magnetic resonance images. arXiv:1801.00415
2. Bahadure NB, Ray AK, Thethi HP (2017) Image analysis for mri based brain tumor detection and feature extraction using biologically inspired bwt and svm. International Journal of Biomedical Imaging, 2017
3. Bahrami K, Shi F, Rekik I, Gao Y, Shen D (2017) 7t-guided super-resolution of 3t mri. Medical Physics 44(5):1661–1677
4. Bakas S, Akbari H, Sotiras A, Bilello M, Rozycki M, Kirby JS, Freymann JB, Farahani K, Davatzikos C (2017) Advancing the cancer genome atlas glioma mri collections with expert segmentation labels and radiomic features. Scientific data 4:170117
5. Bakas S, Reyes M, Jakab A, Bauer S, Rempfler M, Crimi A, Shinozaki RT, Berger C, Hu SM, Rozycki M et al (2018) Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the brats challenge. arXiv:1811.02629
6. Benou A, Veksler R, Friedman A, Raviv TR (2017) Ensemble of expert deep neural networks for spatio-temporal denoising of contrast-enhanced mri sequences. Medical Image Analysis 42:145–159
7. Bergeest JP, Jäger F (2008) A comparison of five methods for signal intensity standardization in mri. In: Bildverarbeitung für die medizin 2008. Springer, pp 36–40
8. Bermudez C, Plassard AJ, Davis LT, Newton AT, Resnick SM, Landman BA (2018) Learning implicit brain mri manifolds with deep learning. In: Medical imaging 2018: Image processing, vol 10574. International Society for Optics and Photonics, p 1057411
9. Betancur J, Commandeur F, Motlagh M, Sharir T, Einstein AJ, Bokhari S, Fish MB, Ruddy TD, Kaufmann P, Sinusas AJ, Miller EJ, Bateman TM, Dorbala S, Di Carli M, Germano G, Otaki Y, Tamarappoo BK, Dey D, Berman DS, Slomka PJ (2018) Deep learning for prediction of obstructive disease from fast myocardial perfusion spect: a multicenter study. JACC Cardiovasc Imaging 11(11):1654–1663. https://doi.org/10.1016/j.jcmg.2018.01.020
10. Betancur J, Hu LH, Commandeur F, Sharir T, Einstein AJ, Fish MB, Ruddy TD, Kaufmann PA, Sinusas AJ, Miller EJ, Bateman TM, Dorbala S, Di Carli M, Germano G, Otaki Y, Liang JX, Tamarappoo BK, Dey D, Berman DS, Slomka PJ (2019) Deep learning analysis of upright-supine high-efficiency spect myocardial perfusion imaging for prediction of obstructive coronary artery disease: a multicenter study. J Nucl Med 60(5):664–670. https://doi.org/10.2967/jnumed.118.213538
11. Chen T, Xiao F, Yu Z, Yuan M, Xu H, Lu L (2021) Detection and grading of gliomas using a novel two-phase machine learning method based on mri images. Front Neurosci, 15
12. Chen W, Liu B, Peng S, Sun J, Qiao X (2018) S3d-unet: separable 3d u-net for brain tumor segmentation. In: International MICCAI brainlesion workshop. Springer, pp 358–368
13. Chen Y, Shi F, Christodoulou AG, Xie Y, Zhou Z, Li D (2018) Efficient and accurate mri super-resolution using a generative adversarial network and 3d multi-level densely connected network. In: International conference on medical image computing and computer-assisted intervention. Springer, pp 91–99
14. Cho HH, Park H (2017) Classification of low-grade and high-grade glioma using multi-modal image radiomics features. In: 2017 39Th annual international conference of the IEEE engineering in medicine and biology society (EMBC). IEEE, pp 3081–3084
15. Çoktuk Ö, Abdulkadir A, Lienkamp SS, Brox T, Ronneberger O (2016) 3d u-net: learning dense volumetric segmentation from sparse annotation. In: International conference on medical image computing and computer-assisted intervention. Springer, pp 424–432
16. Cihat-Er F, Firat Z, Kovanlikaya I, Ture U, Ozurt-Isik E (2018) Machine-learning in grading of gliomas based on multi-parametric magnetic resonance imaging at 3t. Comput ssBiol Med 99:154–160
17. Coupet M, Urruty T, Leelanupah T, Naudin M, Bourdon P, Fernandez-Maloigne C, Guillemin R (2020) An empirical study of deep neural networks for glioma detection from mri sequences. In: International conference on neural information processing. Springer, pp 113–125
18. Dequidt P, Bourdon P, Ahmed OB, Tremblais B, Guillemin C, Naudin M, Fernandez-Maloigne C, Guillemin R (2019) Recent advances in glioma grade classification using machine and deep learning on mr data. In: 2019 Fifth international conference on advances in biomedical engineering (ICABME). IEEE, pp 1–4
19. Dequidt P, Bourdon P, Tremblais B, Guillemin C, Gianelli B, Boutet C, Cottier JP, Vallée JN, Fernandez-Maloigne C, Guillemin R (2021) Exploring radiologic criteria for glioma grade classification on the brats dataset. IRBM
20. Falk T, Mai D, Bensch R, Çiçek Ö, Abdulkadir A, Marrakchi Y, Böhm A, Deubner J, Jückel Z, Seiwald K et al (2019) U-net: deep learning for cell counting, detection, and morphometry. Nature Methods 16(1):67–70
21. Feng Y, Pan H, Meyer C, Feng X (2018) A self-adaptive network for multiple sclerosis lesion segmentation from multi-contrast mri with various imaging protocols. arXiv:1811.07491
22. Ge C, Gu IYH, Jakola AS, Yang J (2018) Deep learning and multi-sensor fusion for glioma classification using multistream 2d convolutional networks. In: 2018 40th annual international conference of the IEEE engineering in medicine and biology society (EMBC). IEEE, pp 5894–5897
23. Goodenberger ML, Jenkins RB (2012) Genetics of adult glioma. Cancer Genetics 205(12):613–621
24. He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp 770–778
25. Huang G, Liu Z, Van Der Maaten L, Weinberger KQ (2017) Densely connected convolutional networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp 4700–4708
26. İşın A, Direkoğlu C, Şah M (2016) Review of mri-based brain tumor image segmentation using deep learning methods. Procedia Comput Sci 102:317–324
27. Jager F, Hornegger J (2008) Nonrigid registration of joint histograms for intensity standardization in magnetic resonance imaging. IEEE Trans Med Imaging 28(1):137–150
28. Kalaiselvi T, Karthiagelsivi S (2015) A novel wavelet based feature selection to classify abnormal images from t2-w axial head scans. In: National conferences on new horizons in computational intelligence and information systems. vol 1, pp 140–145
29. Kalaiselvi T, Padmapriya T, Srimakrishnan P, Priyadharshini V (2020) Development of automatic glioma brain tumor detection system using deep convolutional neural networks. Int J Imaging Syst Technol 30(4):926–938
30. Kaufmann J, Müller K, Montavon G (2020) Towards explaining anomalies: a deep taylor decomposition of one-class models. Pattern Recognit 101:107198. https://doi.org/10.1016/j.patcog.2020.107198
31. Khalid NEA, Ibrahim S, Haniff P (2011) Mri brain abnormalities segmentation using k-nearest neighbors(k-nn). Int J Comput Sci Eng 3(2):980–990
32. Khan MA, Lali IU, Rehman A, Ishaq M, Saba T, Zahoor S, Akram T (2019) Brain tumor detection and classification: a framework of marker-based watershed algorithm and multilevel priority features selection. Microscopy Research and Technique 82(6):909–922
33. Khawaldeh S, Pervaiz U, Rafiq A, Alkhawaldeh RS (2018) Noninvasive grading of glioma tumor using magnetic resonance imaging with convolutional neural networks. Appl Sci 8(1):27
34. Kline TL, Korfiatis P, Edwards ME, Blais JD, Czerwicz FS, Harris PC, King BF, Torres VE, Erickson BJ (2017) Performance of an artificial multi-observer deep neural network for fully automated segmentation of polycystic kidneys. J Digital Imaging 30(4):442–448
35. Korfiatis P, Kline TL, Lachance DH, Parney IF, Buckner JC, Erickson BJ (2017) Residual deep convolutional neural network predicts mgmt methylation status. J Digital Imaging 30(5):622–628
36. Lapuschkin S, Wäldchen S, Binder A, Montavon G, Samek W, Müller KR (2019) Unmasking clever hans predictors and assessing what machines really learn. Nature Communications 10(1):1–8
37. LeCun Y, Bengio Y, Hinton G (2015) Deep learning. Nature 521(7553):436
38. Li H, Parikh NA, He L (2018) A novel transfer learning approach to enhance deep neural network classification of brain functional connectomes. Frontiers in Neuroscience 12:491
39. Li Y, Shen L (2018) Skin lesion analysis towards melanoma detection using deep learning network. Sensors 18(2):556
40. Litjens G, Kooi T, Bejnordi BE, Setio AAA, Ciompi F, Ghafoorian M, Van Der Laak JA, Van Ginneken B, Sánchez CI (2017) A survey on deep learning in medical image analysis. Medical Image Analysis 42:60–88
41. Makropoulos A, Counsell SJ, Rueckert D (2018) A review on automatic fetal and neonatal brain mri segmentation. NeuroImage 173:231–248
42. Menze BH, Jakab A, Bauer S, Kalpathy-Cramer J, Farahani K, Kirby J, Burren Y, Porz N, Slotboom J, Wiest R et al (2014) The multimodal brain tumor image segmentation benchmark (brats). IEEE Transactions on Medical Imaging 34(10):1993–2024
43. Montavon G, Binder A, Lapuschkin S, Samek W, Müller K (2019) Layer-wise relevance propagation: An overview. In: Samek W, Montavon G, Vedaldi A, Hansen LH, Müller K (eds) Explainable AI: Interpreting, Explaining and Visualizing Deep Learning, Lecture Notes in Computer Science, vol 11700. Springer, pp 193–209. https://doi.org/10.1007/978-3-030-28954-6_10
44. Narin A, Kayar C, Pamuk Z (2020) Automatic detection of coronavirus disease (covid-19) using x-ray images and deep convolutional neural networks. arXiv:2003.10849
46. Oktay O, Schlemper J, Folgoc LL, Lee M, Heinrich M, Misawa K, Mori K, McDonagh S, Hammerla NY, Kainz B et al (2018) Attention u-net: Learning where to look for the pancreas. arXiv:1804.03999
47. Pereira S, Pinto A, Alves V, Silva CA (2015) Deep convolutional neural networks for the segmentation of gliomas in multi-sequence mri. In: Brainles 2015. Springer, pp 131–143
48. Ramon AJ, Yang Y, Pretorius PH, Johnson KL, King MA, Wernick MN (2018) Initial investigation of low-dose spect-npi via deep learning. In: 2018 IEEE Nuclear Science Symposium and Medical Imaging Conference Proceedings (NSS/MIC). pp 1–3. https://doi.org/10.1109/NSSMIC.2018.8824548
49. Ronneberger O, Fischer P, Brox T (2015) U-net: Convolutional networks for biomedical image segmentation. In: International conference on medical image computing and computer-assisted intervention. Springer, pp 234–241
50. Roser M, Ritchie H (2020) Cancer our world in data. https://ourworldindata.org/cancer
51. Russakovsky O, Deng J, Su H, Krause J, Satheesh S, Ma S, Huang Z, Karpathy A, Khosla A, Bernstein M, Berg AC, Fei-Fei L (2015) Imagenet large scale visual recognition challenge. International Journal of Computer Vision (IJCV) 115(3):211–252. https://doi.org/10.1007/s11263-015-0816-y
52. Samek W (2019) Explainable AI: interpreting, explaining and visualizing deep learning, vol 11700. Springer Nature, Berlin
53. Schmidhuber J (2015) Deep learning in neural networks: an overview. Neural Networks 61:85–117
54. Shahzadi I, Tang TB, Meriaudeau F, Quyuum A (2018) Cnn-lstm: Cascaded framework for brain tumour classification. In: 2018 IEEE-EMBS Conference on biomedical engineering and sciences (IECBES). IEEE, pp 633–637
55. Simonyan K, Zisserman A (2014) Very deep convolutional networks for large-scale image recognition. arXiv:1409.1556
56. Springenberg JT, Dosovitskiy A, Brox T, Riedmiller MA (2015) Striving for simplicity: The all convolutional net. In: Bengio Y, LeCun Y (eds) 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Workshop Track Proceedings. arXiv:1412.6806
57. Sun J, Li H, Xu Z et al (2016) Deep admn-net for compressive sensing mri. In: Advances in neural information processing systems. pp 10–18
58. Sun X, Shi L, Luo Y, Yang W, Li H, Liang P, Li K, Mok VC, Chu WC, Wang D (2015) Histogram-based normalization technique on human brain magnetic resonance images from different acquisitions. Biomedical Engineering Online 14(1):1–17
59. Szegedy C, Vanhoucke V, Ioffe S, Shlens J, Wojna Z (2016) Rethinking the inception architecture for computer vision. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp 2818–2826
60. Tajbakhsh N, Shin JY, Gurudru SR, Hurst RT, Kendall CB, Gotway MB, Liang J (2017) On the necessity of fine-tuned convolutional neural networks for medical imaging. In: Lu L, Zheng Y, Carneiro G, Yang L (eds) Deep Learning and Convolutional Neural Networks for Medical Image Computing - Precision Medicine, High Performance and Large-Scale Datasets. Advances in Computer Vision and Pattern Recognition, Springer, pp 181–193. https://doi.org/10.1007/978-3-319-42999-1_11
61. Tan M, Le Q (2019) EfficientNet: Rethinking model scaling for convolutional neural networks. In: Chaudhuri K, Salakhutdinov R (eds) Proceedings of the 36th international conference on machine learning, proceedings of machine learning research, vol 97. PMLR, pp 6015–6014. http://proceedings.mlr.press/v97/tan19a.html
62. Urbanska K, Sokolowska J, Szmidt M, Sysa P (2014) Glioblastoma multiforme—an overview. Contemporary oncology 18(5):307
63. Wang S, Su Z, Ying L, Peng X, Zhu S, Liang F, Feng D, Liang D (2016) Accelerating magnetic resonance imaging via deep learning. In: 2016 IEEE 13th international symposium on biomedical imaging (ISBI). IEEE, pp 514–517
64. Wason R (2018) Deep learning: Evolution and expansion. Cogn Syst Res 52:701–708
65. Zhou X, Wang S, Xu W, Ji G, Phillips P, Sun P, Zhang Y (2015) Detection of pathological brain in mri scanning based on wavelet-entropy and na"ive bayes classifier. In: International conference on bioinformatics and biomedical engineering. Springer, pp 201–209

Publisher’s note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.