Binary Glioma Grading: Radiomics versus Pre-trained CNN Features

Milan Decuyper(✉), Stijn Bonte, and Roel Van Holen

Medical Imaging and Signal Processing, Ghent University, Ghent, Belgium
milan.decuyper@ugent.be

Abstract. Determining the malignancy of glioma is highly important for initial therapy planning. In current clinical practice, often a biopsy is performed to verify tumour grade which involves risks and can negatively impact overall survival. To avoid biopsy, non-invasive tumour characterisation based on MRI is preferred and to improve accuracy and efficiency, the use of computer-aided diagnosis (CAD) systems is investigated. Existing radiomics CAD techniques often rely on manual segmentation and are trained and evaluated on data from one clinical centre. Therefore, there is a need for accurate and automatic CAD systems that are robust to large variations in imaging protocols between different institutions. In this study, we extract features from T1ce MRI with a pre-trained CNN and compare their predictive power with hand-engineered radiomics features for binary grade prediction. Performance was evaluated on the BRATS 2017 database containing MRI and manual segmentation data of 285 patients from multiple institutions. State-of-the-art performance with an AUC of 96.4% was achieved with radiomics features extracted from manually segmented tumour volumes. Pre-trained CNN features had a strong predictive value as well and an AUC score of 93.5% could be obtained when propagating the tumour region of interest (ROI). Additionally, using a pre-trained CNN as feature extractor, we were able to design an accurate, automatic, fast and robust binary glioma grading system achieving an AUC score of 91.1% without requiring ROI annotations.

1 Introduction

The optimal treatment strategy of newly diagnosed glioma strongly relies on tumour malignancy. Diffuse glioma, the most common form of primary brain tumours, are divided into grades II to IV according to malignancy by the World Health Organization (WHO) [1]. Glioblastoma multiform (GBM) is the most aggressive type of primary brain tumour and has a very poor prognosis with a 5-year survival rate of only 4–5% [2]. Current standard of care for GBMs consists of early resection combined with chemotherapy and radiotherapy. Lower-grade gliomas (LGGs), on the other hand, have more favourable outcomes and possible treatment strategies include: a wait-and-scan approach, a biopsy for histopathological verification or immediate resection [3]. A recent study by Wijnenga et

A. F. Frangi et al. (Eds.): MICCAI 2018, LNCS 11072, pp. 498–505, 2018.
https://doi.org/10.1007/978-3-030-00931-1_57

al. [3] shows that biopsy as initial strategy negatively impacts overall survival with a reported hazard ratio of 2.69 (95% CI 1.19–6.06; $p = 0.02$) compared to wait-and-scan. The invasive procedure involves high risks, is subject to sampling error and the results may be subjective, depending on the neuropathologist performing the histopathological analysis [4]. Hence a biopsy to confirm diagnosis and grade of the tumour should be avoided and accurate non-invasive grading is preferred.

Conventional MR imaging with gadolinium-based contrast agents is an established technique for non-invasive brain tumour characterisation [5,6]. Through MRI, information is obtained regarding contrast enhancement, necrosis, oedema, mass effect, which are considered important predictors of tumour malignancy. Nevertheless, brain tumour grading using this diagnostic technique is not always reliable with reported sensitivities ranging between 55% and 83% [5]. For example, low-grade glioma demonstrating contrast enhancement can be misdiagnosed as high-grade or conversely 40–45% of non-enhancing lesions are found to be highly malignant gliomas after histopathological verification [6]. Moreover, the ever-increasing amount of MR image data raises the burden of accurate data analysis and dramatically increases the workload of radiologists.

Computer-aided diagnosis (CAD) may provide a way to handle this data explosion and increase diagnostic accuracy [7]. CAD systems can automatically process MR images, calculate quantitative features describing tumour characteristics and combine them to estimate tumour type and grade through the use of artificial intelligence. The time required for diagnosis can be reduced and accuracy and treatment planning enhanced while avoiding the need for biopsy. Towards computer-aided brain tumour diagnosis, the use of radiomics has been investigated [7–9]. Radiomics involves the extraction and analysis of quantitative image features and typically consists of three stages: tumour segmentation, feature extraction and finally classification or analysis of the radiomics features. Zacharaki et al. [8] investigated the classification of brain tumours into different types and grades based on conventional and perfusion MRI. In the proposed method, shape, intensity and Gabor texture features were extracted from regions of interest manually traced by expert neuroradiologists. On a dataset of 102 glioma from 98 patients, an accuracy of 87% was achieved for discriminating high-grade from low-grade glioma with a support vector machine (SVM). A system for grade identification (low- versus high-grade) of astrocytoma from T2-weighted images was designed in the work by Subashini et al. [9]. Tumours were isolated with fuzzy c-means segmentation from which shape, intensity and texture features were calculated. A learning vector quantisation classifier trained on 164 images and evaluated on 36 images achieved an accuracy of 91%. An overview of MRI based medical image analysis studies regarding brain tumour segmentation and grade classification is provided by Mohan and Subashini [7]. In current radiomics studies, often input of domain experts is required, such as manual segmentation data, making these methods not reproducible and not fully automatic. Additionally, most CAD methods are trained and evaluated on data from one clinical centre. Hence these systems are potentially not robust or applicable to data from other centres due to large variations in imaging protocols.

Our goal is to investigate the use of deep learning to develop an accurate, reproducible and fully automatic CAD system. State-of-the-art deep learning models, like convolutional neural networks (CNNs) achieve high performances in object recognition tasks [10]. We investigate the application of these techniques on medical imaging data and study their performance for brain tumour diagnosis. Deep learning has extensively been used in medical image analysis [11] and is increasingly employed in brain tumour segmentation challenges [12]. Binary brain tumour grading using a CNN trained from scratch on data from BRATS 2014 was evaluated by Pan et al. [13]. Sensitivity and specificity scores of 73% were achieved with only a limited and imbalanced dataset. Automated diagnosis with deep learning remains a challenging task as large-scale datasets of brain tumour scans comparable to ImageNet are unavailable. Therefore, in this work, we will try to overcome this lack of large training sets through the use of transfer learning. The application of pre-trained CNNs for survival prediction based on MRI has been investigated by Ahmed et al. [14]. An accuracy of 82% was achieved for differentiating long-term from short-term survival cases on a limited dataset of 22 GBM patients.

To conclude, state-of-the-art performance in binary tumour grading is currently achieved through radiomics with reported accuracies of 87% up to 91%. Only one study using deep learning for binary grade prediction was found reaching sensitivity and specificity scores of 73%. In this paper, we investigate the use of hand-engineered radiomics features and features extracted through a pre-trained CNN to achieve state-of-the-art performance in discriminating GBMs from lower-grade glioma. This allowed us to compare the predictive value of the radiomics features with pre-trained CNN features on the same heterogeneous dataset. In the radiomics approach, shape, intensity and texture features are extracted from T1ce scans manually segmented into different tumour tissues. Deep features, on the other hand, are extracted using a CNN trained on ImageNet [10].

2 Materials and Methods

2.1 Data

The data used in this work originates from the BRATS 2017 database [12,15]. It contains multi-institutional routine clinically-acquired pre-operative MRI scans of 210 glioblastoma (GBMs) and 75 lower-grade glioma (WHO grade II and III) with pathologically confirmed diagnosis. For each case a T1, T2, T1ce and FLAIR sequence is available. The MRI scans originate from multiple institutions and were acquired with different clinical protocols and scanners resulting in a very heterogeneous dataset. All subject's sequences are co-registered to the same anatomical template, interpolated to a 1 mm^3 voxel size and skull-stripped. Additionally, manual segmentation labels are provided denoting the GD-enhancing, peritumoural oedema and the necrotic and non-enhancing tumour regions. In this study, only the T1ce sequence and segmentation data were used to perform binary grade prediction.

2.2 Feature Extraction: Radiomics

In the radiomics feature extraction approach, all scans were first bias corrected using SPM12 (version 6906, Wellcome Trust Centre for Neuroimaging, University College London) running on MATLAB R2017b (The MathWorks, Inc., Natick, MA). Next, since MRI scans are recorded in arbitrary units, the image intensities were normalised following the robust white stripe normalisation [16]. The manual segmentation labels were used to define five different tumour regions: total abnormal region, tumour core, enhancing tissue, necrosis and oedema. In every region we calculated 207 quantitative features: 14 histogram, 8 size and shape, 138 grey-level co-occurence, 22 grey-level run-length matrix, 12 neighbourhoord grey-tone difference matrix and 13 grey-level size-zone matrix features, according to the definitions in Aerts et al. [17] and Willaime et al. [18].

2.3 Feature Extraction: Pre-trained CNN

Instead of extracting hand-engineered features from the segmented tumour volumes, deep features were extracted using a pre-trained convolutional neural network. The VGG-11 architecture was used consisting of 8 convolutional and 3 fully connected layers [19]. The model, pre-trained on the ImageNet dataset, was loaded from the pyTorch torchvision package. Features were obtained by forward propagating an MRI slice through the network and extracting the 4096-dimensional output of the first fully connected layer. The first layer was chosen under the assumption that earlier layers learn more generally applicable features than layers deeper into the network. Before being propagated through the network, the slices were pre-processed to match the expected input of the pre-trained pytorch models. The image intensities were scaled to a range between [0,1], the slice was resized to a shape of 224×224 through bilinear interpolation and finally normalised with mean and standard deviation values provided by pyTorch. Because the model expects RGB images, the MRI slice was provided at the R channel and the B and G channels were set to zero.

Feature extraction and corresponding grading performance was evaluated for four different ways of providing the T1ce scan at the input of the network (see Fig. 1). In a first approach, the segmentation data was used to select the slice in the T1ce scan containing the largest tumour contour and crop this slice to the size of the tumour (Fig. 1: method 1). After applying the pre-processing steps explained above, the tumour patch was propagated through the network, thereby obtaining one 4096-dimensional feature vector with a corresponding label indicating LGG or GBM.

For the second method, all tumour slices were propagated through the network after being cropped to the size of the tumour (Fig. 1: method 2). Hence, multiple feature vectors are obtained for each patient and every slice or feature vector was classified into one of three classes: (1) LGG, (2) GBM where only oedema is visible, (3) GBM with contrast enhancement and necrosis. In each slice, either a LGG or a GBM is visible. Additionally, a GBM may in some slices only display oedema and no contrast enhancement and necrosis. Because these slices may have a similar appearance as LGG slices, this could be confusing for the classifier and therefore a separate class was added for GBM slices only demonstrating oedema.

In the third method, the same slice was selected as in the first approach, but now it was not cropped (Fig. 1: method 3). Hence the entire slice was propagated through the network.

To design a system able to classify a T1ce scan without requiring segmentation information, a fourth method was investigated. Here, every slice of the T1ce scan was propagated through the network (Fig. 1: method 4). One entire scan contains 155 slices, so 155 feature vectors were obtained for each patient and a fourth class, besides the three classes of the second method, was added for slices containing no tumour. Using this approach, no segmentation data is required to classify slices from a T1ce sequence of a new patient resulting in a fully automatic CAD system.

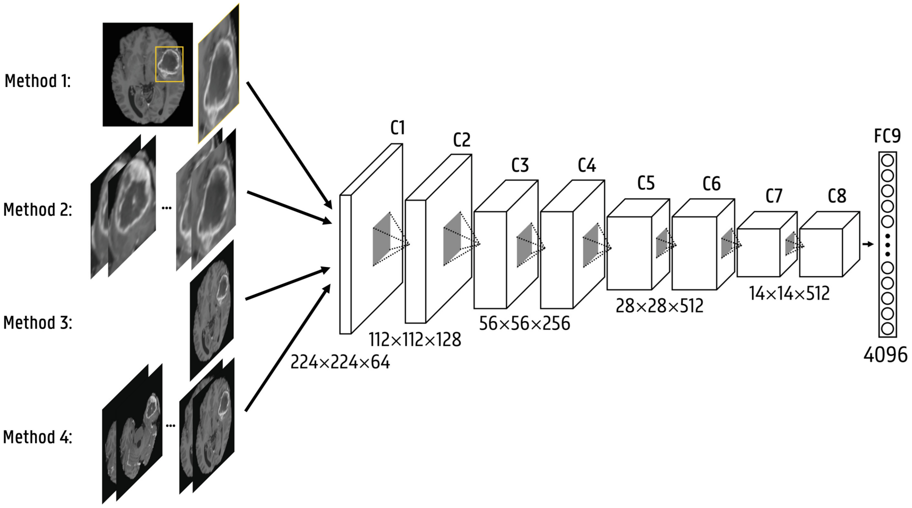

Fig. 1. Feature extraction with the pre-trained VGG-11 CNN. Method 1: Propagate tumour region of the slice containing the largest tumour contour. Method 2: Propagate tumour region of all tumour slices. Method 3: Propagate entire slice containing the largest tumour contour. Method 4: Propagate all slices

2.4 Classification

After feature extraction, classification was performed with the goal to predict whether a patient has a glioblastoma or lower-grade glioma. The feature vectors were first scaled to unit norm and features showing no variance between different samples were removed. For classification, the python scikit-learn *RandomForestClassifier* was used with 200 decision trees. All Random Forest (RF) models were trained for the binary classification task except for the second and fourth method of feature extraction with the pre-trained CNN. In those cases, the RF model was trained to classify a slice into one of 3, respectively 4 classes as explained in Sect. 2.3. For each patient, multiple slices were classified. All predictions were combined by calculating their mean probability and the sum of the probabilities of the two GBM classes was used as the final probability value of having a GBM. The performance of the classifier was evaluated on a separate test set containing 57 (20%) of the 285 glioma cases. The class ratio of 210:75 was equal in both training and test set. To enhance sensitivity and specificity of the model, the probability threshold of classifying a glioma as GBM was optimised through 5-fold cross-validation. The training and evaluation process was repeated 50 times with different random splits in train and test set to estimate average performance and variability of the model.

3 Results

For each of the feature extraction methods, a RF model was trained and evaluated to asses the predictive value of the resulting feature vectors. The area under the ROC curve (AUC), accuracy, sensitivity and specificity scores are reported in Table 1. The RF model trained on the radiomics features achieves the highest performance with an average AUC score of 96%. With features extracted using a pre-trained CNN, best results were obtained when zooming in on the tumour region and using all tumour slices (CNN, method 2). When using features extracted from the entire slice containing the largest tumour contour (CNN, method 3), performance is lower with an AUC of 87% compared to 92%. However, when predicting glioma grade based on all slices of the T1ce scan (CNN, method 4), performance could be improved to an AUC score of 91%. Classifying a T1ce scan was possible within 0.3 s with *CNN: method 1* and *3*, 12 s with *CNN: method 2* and 30 s with *CNN: method 4* on a Macbook Pro with 2.8 GHz Intel Core i7 CPU where propagating all slices through the CNN required most of the computation time.

Table 1. Mean (std) (%) area under the ROC curve, accuracy, sensitivity and specificity classification scores.

Feature extraction method	AUC	Acc.	Sens.	Spec.
Radiomics	96.4(2.6)	89.6(3.8)	89.9(5.4)	88.8(8.6)
CNN: Method 1	92.2(3.9)	83.8(4.6)	83.3(5.2)	85.2(9.6)
CNN: Method 2	93.5(3.0)	86.1(4.3)	85.4(5.4)	88.5(8.1)
CNN: Method 3	86.8(4.6)	79.1(4.9)	78.6(6.4)	80.7(9.6)
CNN: Method 4	91.1(3.6)	82(5.3)	81.5(7.2)	83(9.6)

4 Discussion

The results shown in Table 1 show that the best performance is achieved with the radiomics approach, matching or even outperforming state-of-the-art accuracies reported today. The achieved performance, however, was obtained when extracting radiomics features from manually segmented tumour tissues which is time-consuming and introduces subjectivity. A lot of research has been performed towards automatic segmentation algorithms and the difference in performance between using a state-of-the-art automatic segmentation algorithm or manual segmentation remains to be investigated.

Although performance is slightly lower compared to the radiomics results, accurate grading could be achieved with a pre-trained CNN as feature extractor as well. With the first method of feature extraction through a CNN, an AUC is achieved of 92% while only requiring a bounding box around the tumour which is considerably less time-consuming than accurate segmentation of the different tissues. Furthermore, when estimating grade based on all tumour slices, performance could be improved to an AUC of 93.5%. These classification scores are more than 10% higher than currently reported binary grading performance with deep learning. Moreover, an automatic segmentation algorithm could be used to define the bounding box and we expect that small variations or inaccuracies will not have a large influence on performance. Features extracted from the entire slice were less informative but by calculating an ensemble prediction from all slices, accurate grading could still be achieved reaching a performance similar to the first method. This way, a binary grading system could be designed that is fast, does not require segmentation or manual input to classify new T1ce sequences and is trained on a very heterogeneous dataset making it robust to variations in imaging protocols. These results show that a CNN, trained on an entirely different image dataset containing natural images, is able to extract informative features from MRI sequences as well. Their predictive value is lower than radiomics features extracted from manually segmented tumour volumes, but we expect that by fine-tuning the network on brain tumour MRI, results can further be improved. Future work will focus on gathering more data, allowing to specialise CNNs on brain MRI and open the path towards more accurate and automatic brain tumour characterisation.

5 Conclusion

In this work, we compared the predictive value of radiomics features with features extracted using a pre-trained CNN for binary brain tumour grading. Classification results showed that the best performance is achieved with shape, intensity and texture features extracted from manually segmented tumour volumes. Features from a pre-trained CNN, on the other hand, had a high predictive value as well and allowed to design an accurate, fast, automatic and robust binary grading system. These results indicate that a pre-trained CNN, with possible fine-tuning and more data, holds the potential to develop an accurate, reproducible an fully automatic CAD system.

References

1. Louis, D.N., Perry, A., et al.: The 2016 world health organization classification of tumors of the central nervous system: a summary. Acta Neuropathol. **131**(6), 803–820 (2016)
2. Carlsson, S.K., Brothers, S.P., Wahlestedt, C.: Emerging treatment strategies for glioblastoma multiforme. EMBO Mol. Med. **6**(11), 1359–1370 (2014)
3. Wijnenga, M.M.J., Mattni, T., et al.: Does early resection of presumed low-grade glioma improve survival? A clinical perspective. J. Neuro-Oncol. **133**(1), 137–146 (2017)
4. Jackson, R.J., Fuller, G.N., et al.: Limitations of stereotactic biopsy in the initial management of gliomas. Neuro-oncology **3**(3), 193–200 (2001)
5. Law, M., Yang, S., et al.: Glioma grading: sensitivity, specificity, and predictive values of perfusion MR imaging and proton MR spectroscopic imaging compared with conventional MR imaging. AJNR **24**(10), 1989–1998 (2003)
6. Jansen, N.L., Graute, V., et al.: MRI-suspected low-grade glioma: is there a need to perform dynamic FET PET? EJNMMI **39**(6), 1021–1029 (2012)
7. Mohan, G., Subashini, M.M.: MRI based medical image analysis: survey on brain tumor grade classification. Biomed. Signal Process. Control. **39**, 139–161 (2018)
8. Zacharaki, E.I., Wang, S., et al.: MRI-based classification of brain tumor type and grade using SVM-RFE. In: Proceedings - 2009 IEEE International Symposium on Biomedical Imaging: From Nano to Macro, ISBI 2009, pp. 1035–1038. IEEE, June 2009
9. Subashini, M.M., Sahoo, S.K., et al.: A non-invasive methodology for the grade identification of astrocytoma using image processing and artificial intelligence techniques. Expert. Syst. Appl. **43**, 186–196 (2016)
10. Russakovsky, O., Deng, J., et al.: ImageNet large scale visual recognition challenge. Int. J. Comput. Vis. **115**(3), 211–252 (2015)
11. Litjens, G., Kooi, T., et al.: A survey on deep learning in medical image analysis. Med. Image Anal. **42**(December 2012), 60–88 (2017)
12. Menze, B.H., Jakab, A., et al.: The multimodal brain tumor image segmentation benchmark (BRATS). IEEE Trans. Med. Imaging **34**(10), 1993–2024 (2015)
13. Pan, Y., Huang, W., et al.: Brain tumor grading based on neural networks and convolutional neural networks. In: 2015 37th Annual International Conference of the EMBC, pp. 699–702. IEEE, August 2015
14. Ahmed, K.B., Hall, L.O., et al.: Fine-tuning convolutional deep features for MRI based brain tumor classification. In: Proceedings of SPIE 10134, Medical Imaging 2017: Computer-Aided Diagnosis, p. 101342E, March 2017
15. Bakas, S., Akbari, H., et al.: Advancing the cancer genome atlas glioma MRI collections with expert segmentation labels and radiomic features. Sci. Data **4**, 170117 (2017)
16. Shinohara, R.T., Sweeney, E.M., et al.: Statistical normalization techniques for magnetic resonance imaging. NeuroImage Clin. **6**, 9–19 (2014)
17. Aerts, H.J.W.L., Velazquez, E.R., et al.: Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. Nat. Commun. **5**, 4006 (2014)
18. Willaime, J.M., Turkheimer, F.E., et al.: Quantification of intra-tumour cell proliferation heterogeneity using imaging descriptors of 18F fluorothymidine-positron emission tomography. Phys. Med. Biol. **58**(2), 187–203 (2013)
19. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. CoRR abs/1409.1, pp. 1–14 (2014)