Radiomics-based convolutional neural network for brain tumor segmentation on multiparametric magnetic resonance imaging

Prateek Prasanna
Ayush Karnawat
Marwa Ismail
Anant Madabhushi
Pallavi Tiwari

Radiomics-based convolutional neural network for brain tumor segmentation on multiparametric magnetic resonance imaging

Prateek Prasanna,^{a,*†} Ayush Karnawat,^{a,†} Marwa Ismail,^a Anant Madabhushi,^{a,b} and Pallavi Tiwari^{a,*}

^aCase Western Reserve University, Department of Biomedical Engineering, Cleveland, Ohio, United States

^bLouis Stokes Cleveland Veterans Administration Medical Center, Cleveland, Ohio, United States

Abstract. Accurate segmentation of gliomas on routine magnetic resonance image (MRI) scans plays an important role in disease diagnosis, prognosis, and patient treatment planning. We present a fully automated approach, radiomics-based convolutional neural network (RadCNN), for segmenting both high- and low-grade gliomas using multimodal MRI volumes (T1c, T2w, and FLAIR). RadCNN incorporates radiomic texture features (i.e., Haralick, Gabor, and Laws) within DeepMedic [a deep 3-D convolutional neural network (CNN) segmentation framework that uses image intensities; a top performing method in the BraTS 2016 challenge] to further augment the performance of brain tumor subcompartment segmentation. We first identify textural radiomic representations that best separate the different subcompartments [enhancing tumor (ET), whole tumor (WT), and tumor core (TC)] on the training set, and then feed these representations as inputs to the CNN classifier for prediction of different subcompartments. We hypothesize that textural radiomic representations of lesion subcompartments will enhance the separation of subcompartment boundaries, and hence providing these features as inputs to the deep CNN, over and above raw intensity values alone, will improve the subcompartment segmentation. Using a training set of $N = 241$ patients, validation set of $N = 44$, and test set of $N = 46$ patients, RadCNN method achieved Dice similarity coefficient (DSC) scores of 0.71, 0.89, and 0.73 for ET, WT, and TC, respectively. Compared to the DeepMedic model, RadCNN showed improvement in DSC scores for both ET and WT and demonstrated comparable results in segmenting the TC. Similarly, smaller Hausdorff distance measures were obtained with RadCNN as compared to the DeepMedic model across all the subcompartments. Following the segmentation of the different subcompartments, we extracted a set of subcompartment specific radiomic descriptors that capture lesion disorder and assessed their ability in separating patients into different survival cohorts (short-, mid- and long-term survival) based on their overall survival from the date of baseline diagnosis. Using a multilinear regression approach, we achieved accuracies of 0.57, 0.63, and 0.45 for the training, validation, and test cases, respectively. © 2019 Society of Photo-Optical Instrumentation Engineers (SPIE) [DOI: [10.1117/1.JMI.6.2.024005](https://doi.org/10.1117/1.JMI.6.2.024005)]

Keywords: gliomas; segmentation; radiomics; feature selection; convolutional neural network.

Paper 18005RR received Jan. 6, 2018; accepted for publication Apr. 15, 2019; published online May 7, 2019.

1 Introduction

Gliomas, one of the most common types of primary brain tumors, exhibit phenotypically heterogeneous subregions comprising of the enhancing and nonenhancing lesion, necrotic core, and the surrounding edema (ED), each of which contains relevant diagnostic and prognostic information.¹ As such, accurate estimation of the information (i.e., volume/position) contained within these regions is critical for diagnosis, treatment planning, and overall patient survival assessment. Toward that end, reliable and accurate delineation of the enhancing tumor (ET) and its subregions [ED, nonenhancing tumor (NET), and necrosis (NCR)] is required, which, due to their variable shape and size, poses a significant challenge. In particular, manual segmentation of different subcompartment boundaries is both time consuming and prone to misinterpretation and human error, often resulting in high inter-rater variability.²

Accurate, automatic segmentation frameworks aim to solve this problem while providing a more efficient and scalable solution for clinical applicability.³ In recent years, convolutional neural networks (CNNs) have drawn increasing attention for problems involving classification and semantic segmentation, especially in the field of image recognition. In fact, leading methods from previous years of the brain tumor segmentation (BraTS) challenge^{4–6} have consistently employed CNN-based architectures. However, these CNN architectures have largely been based on image intensities alone and, in most cases, have large training computational overheads due to relatively small training samples. More recently, studies^{7,8} have employed Gabor filtering as a preprocessing step for training CNNs,⁹ to improve the model's learning effectiveness toward image classification. In these works, Gabor filters are sought to modulate the learnable convolution filters, and thus, reduce the number of learnable network parameters to enhance the robustness of learned features to scale and orientation changes. There is similarly an opportunity to combine higher order radiomic features (i.e., Haralick, Laws features) within the deep CNN network as a

*Address all correspondence to Prateek Prasanna, E-mail: prateek@case.edu; Pallavi Tiwari, E-mail: pvt130@case.edu

†These authors contributed equally to this work.

prior preprocessing step, to potentially augment the CNN model by capturing morphologic attributes that could accentuate the lesion boundaries. Radiomic textural features allow for the capture of higher order quantitative measurements (e.g., co-occurrence matrix homogeneity, neighboring gray-level dependence matrix, multiscale Gaussian derivatives), for modeling macro- and microscale morphologic attributes within and around the lesion area from across different MRI protocols. Texture analysis has previously shown promise in distinguishing different grades of brain tumors¹⁰ and identifying brain tumors from treatment confounders.¹¹ In fact, texture features have previously been used in conjunction with random forests¹² to segment brain tumor lesions. In the first part of the BraTS challenge concerning segmentation, we hypothesize that textural radiomic representations will enhance the separation of subcompartment boundaries, and hence providing these features as inputs to the deep CNN, over and above raw intensity values alone, will improve the subcompartment segmentation.

Specifically, we use hand-crafted features as a preprocessing step outside of the CNN model to allow identification of the optimal set of alternate representations of the original data, and then feed these features as inputs to the CNN to help better segment the brain tumors. Our approach is similar to the ones presented in previous studies.^{13,14} For example, Rawat et al.,¹³ where hand-crafted features (position, shape, and orientation descriptors) from digital pathology slides, were learned using the training set and then fed into a deep CNN to learn spatial patterns that correlate to ER-positive or ER-negative status. Using such an approach with preselected filter responses as inputs could potentially result in compact and more efficient networks, with fewer parameters to be learned.

Another equally challenging problem in glioblastoma multi-forme management is stratifying patients based on their overall survival (OS) on baseline MRI scans. Despite aggressive treatment including maximal surgical resection and chemoradiation therapy,¹⁵ the median survival after diagnosis for high-grade gliomas (HGGs) is only 14 months.¹⁶ With monoclonal antibodies, vaccines, and gene therapies currently under investigation,^{17,18} accurate stratification of survival and risk categories may serve as an important precursor for designing personalized approaches in glioma management. Our team has recently developed a radiomic feature descriptor, named “Co-occurrence of Local Anisotropic Gradient Orientation” (CoLIAGe),¹¹ that captures lesion disorder, and has been shown to differentiate benign radiation changes from tumor recurrence on routinely acquired MRI scans in brain tumor patients. In the second part of the BraTS challenge concerning predicting patient outcome, we hypothesize that local disorder-based CoLIAGe features obtained from routine MRI scans express differentially across long-term, mid-term, and short-term survivors, and a prognostic model using CoLIAGe features can stratify patients into their respective survival cohorts.

Specifically, in this paper, we (1) present radiomics-based CNN (RadCNN), a multiscale CNN architecture that incorporates optimized radiomic texture features as an input to a deep CNN model, for improved estimation of tumor subcompartments: ET, NET, NCR, and ED; and (2) employ a radiomics descriptor, CoLIAGe, extracted from the different tumor subcompartments, and assess their prognostic ability in classifying patients into different survival groups. In the following sections, we illustrate our approach and describe, in detail, the features of our architecture followed by its application in segmentation of gliomas and corresponding survival prediction.

2 Methodology

2.1 Notation

We define an image scene $\mathcal{I}$ as $\mathcal{I} = (C, f)$, where $\mathcal{I}$ is a spatial grid C of voxels $c \in C$, in a three-dimensional (3-D) space, $\mathbb{R}^3$. Each voxel, $c \in C$, is associated with an intensity value $f(c)$. $\mathcal{I}_{ET}$, $\mathcal{I}_{ED}$, and $\mathcal{I}_{NEC}$ correspond to the intratumoral, peritumoral ED, and necrotic core subvolumes within every $\mathcal{I}$, respectively, such that $[\mathcal{I}_{ET}, \mathcal{I}_{ED}, \mathcal{I}_{NEC}] \subset \mathcal{I}$. We define each predicted segmentation map $\mathcal{P}$ obtained from the RadCNN pipeline as an undirected graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$. We let $\mathcal{V} = \{v_1, v_2, v_3, \dots, v_m\}$ be the set of voxels that are either labeled as NCR + NET, ET, or ED, where m is the total number of voxels that satisfy this criterion. The edges $\mathcal{E}$ for postprocessing using connected component analysis and Markov random fields (MRFs) are further defined in Sec. 2.6.

2.2 Workflow

Figure 1 shows the workflow of our presented RadCNN model. Briefly, textural radiomic features are first extracted from the multiparametric MRI scans. This is followed by a selection of the most relevant features for differentiating the various subcompartments. Finally, the best features, in addition to the raw intensity channels, are provided as input to a multiresolution CNN for multiclass classification. The predictions are then refined using connected component analysis, MRF, or, in some cases, both, for use in patient survival analysis. For the survival analysis, we first extract CoLIAGe features from the different subcompartments (obtained from the segmentation step), retain the most relevant features, and use a multilinear regression algorithm to predict the survival time.

2.3 Radiomic Feature Extraction

For each 3-D MRI volume, we chose to primarily extract textural radiomic features, including Gabor,¹⁹ Laws,²⁰ and Haralick²¹ descriptors, due to their ability to capture micro- and macrolevel morphologic attributes relating to intensity, edges, and gradient-specific differences across different compartments on routine MRI scans (T1c, T2w, and FLAIR). A Gabor filter can be defined as the modulation of a complex sinusoidal by a Gaussian function and is controlled by scale λ and orientation θ parameters. Gabor features, which are modeled according to human visual perception, are extracted as a response to the convolution of an image with distinct Gabor filters obtained by varying each of the associated parameters across the filter bank. Haralick features capture gray-level co-occurrence patterns, where a matrix of co-occurring gray-level pairs in the image is constructed, from which second-order statistical texture features can be derived. Second order intensity statistics, such as angular second moment, contrast, and difference entropy, are used to characterize the MRI images. Laws features use 5×5 separable masks that are symmetric or antisymmetric to extract level (L), edge (E), spot (S), wave (W), and ripple (R) patterns on an image. The convolution of these masks with every image yields distinct Laws features. In particular, we computed 40 Gabor filter responses with varying $\lambda = 2, 4, 8, 16, 32$ and $\theta = 0^\circ, 22.5^\circ, 45^\circ, 67.5^\circ, 90^\circ, 112.5^\circ, 135^\circ, 157.5^\circ$ values, 13 Haralick features, and 25 Laws features. In total, we extracted a total of 78 radiomic features for each sequence (T1w, T2, and FLAIR), resulting in a total of 234 texture features per study.

Fig. 1 RadCNN pipeline—workflow of our presented framework, which comprises two stages. First, we extract and select the top radiomic features that best distinguish different compartments via feature selection within a random forest classifier. We then use the selected texture map volumes, along with the normalized multiparametric MRI (T1c, T2w, and FLAIR) scans as channel inputs to a 3-D CNN for classification of the tumor habitat into background (gray), ET (cyan), necrotic core (green), and peritumoral ED zones (yellow). We then extract local disorder-based CoLIAGe features from the annotated subcompartments and predict OS using a multilinear regression model.

2.4 Feature Selection

In order to create a well-defined discriminative model, our goal was to select only the features that would improve the segmentation of the subcompartments that were getting over- or under-segmented using an intensity-based CNN (i.e., DeepMedic model). Hence, our feature selection experiments were driven by the output of the DeepMedic model during the training stage. We observed the following consistent trends while training the cases with the DeepMedic model: (a) oversegmentation of ED regions and (b) undersegmentation of NCR + NET regions. We hence designed two separate feature selection experiments to address the aforementioned problems, where our primary goal was to identify the best texture features to (a) distinguish ED from background (nontumor brain tissue) voxels and (b) distinguish NCR + NET regions from ED and background voxels. We used a minimum redundancy maximum relevance²² algorithm in conjunction with a random forest classifier and evaluated the importance of each of the 234 texture features in a threefold cross-validation setting on a randomly identified subset of voxels from the training dataset.

In each of the 100 iterations, we assigned weights to the selected features based on their rank of occurrence over the cross-validation runs. The cumulative weights over 100 runs were then used to select the best features. The top two features for each classification experiment were retained for use in the CNN.

2.5 Three-Dimensional Convolutional Neural Network

The CNN portion of our pipeline is based on the DeepMedic framework by Kamnitsas et al.,⁴ which has been shown to provide the best-performing automated brain subcompartment segmentations in BraTS'16 challenge benchmark datasets. The

network consists of an 11-layer deep multiscale 3-D CNN consisting of two parallel convolutional pathways that process the input at both a normal resolution and one at a lower scale to achieve a large receptive field for classification and segmentation.⁴

The CNN architecture, as shown in Fig. 1, comprises 11 layers: eight consecutive convolutional-pooling layers followed by two fully connected layers and one classification layer along two pathways. Each convolutional-pooling layer uses the same fixed $3 \times 3 \times 3$ convolutional kernel and $2 \times 2 \times 2$ pooling kernel with 30, 30, 40, 40, 40, 40, 50, and 50 neurons, respectively. Both fully connected layers have 150 neurons, which are connected to the four final neurons to determine each voxel's region subtype. We utilized Adam optimizer, which has been shown to work well in practice and compares favorably to other adaptive learning-method algorithms.²³ Most importantly, our network includes additional input channels in the form of selected textural radiomic maps extracted from the original multiparametric MRI scans.

To measure the value of preselecting certain low-level hand-crafted features as input to a deep 3-D CNN framework, we quantitatively compared the performance of our model against a standard DeepMedic model trained on multiparametric intensities by modifying the number of input channels into the network.

2.6 Postprocessing

Although RadCNN was able to identify different tumor subcompartments with high accuracy, we noticed a few issues with the results, namely (a) there were small regions of ED predicted outside and away from the main tumor region and (b) some boundary voxels between the ET and NCR + NET regions were incorrectly predicted as normal tissue (i.e., background) when compared to ground-truth (GT). In order to further refine our segmentation and improve results, we utilized connected component analysis²⁴ and an MRF²⁵ approach on the predictions.

2.6.1 Connected component analysis

When addressing the issue of disconnected segmentations, we wanted to ensure that we kept the segmentation of the main lesion and its subcompartments intact while removing disconnected regions that were not part of the main tumor. We first apply a Gaussian kernel with standard deviation $\sigma = 2$ to each predicted segmentation map $\mathcal{P}$ obtained from the RadCNN pipeline. This smooths out the predictions, essentially combining smaller regions to form bigger connected components. We then create a mask of the regions, where each voxel value $f(c) < \frac{1}{v} \sum_{i=1}^v f(c_i)$, that is, voxels where the value at a specific location c is less than the mean value of all voxels within that image, and accordingly update our prediction map.

We define the edges $\mathcal{E}$ that connect the vertices $\mathcal{V}$ together by performing a depth-first search, creating a set of connected components regions $\mathcal{R} = \{r_1, r_2, \dots, r_t\}$, where t is the total number of regions. For each region r_i , we count the number of vertices (aka voxels) within that region and remove those with less than h vertices, where h represents the mean number of vertices for all regions. This leaves us with the main lesion as well as its subcompartments with far away disconnected segmentations removed.

2.6.2 Markov random field

To address the issue of incomplete segmentations in-between different tumor subcompartments, we predicted the likelihood of voxels belonging to one of the three tumor subcompartments, based on previously predicted segmentations. This allowed us to predict the category of each voxel, thereby improving segmentation results.

We first define the edges $\mathcal{E}$ (neighbors) that connect vertices $\mathcal{V}$ together as voxels that can have at most a fifth-order neighborhood system. In other words, the distance between the current voxel v and another voxel can be at most 5 voxels in any of the X, Y, Z directions. Note that normal brain tissue is not considered as a part of neighborhood voxels. Thus, boundary voxels will have fewer neighbors than interior voxels since boundary voxels will border more normal brain tissue. Using this $\mathcal{G}$, we apply a standard MRF model on our prediction map $\mathcal{P}$ to aggregate the predictions on a voxel level.

2.7 Survival Analysis

For the survival analysis, we extracted 3-D gradient-based CoLIAGE descriptors¹¹ that capture lesion heterogeneity from regions of (a) ET, (b) the peritumoral edematous zone, and (c) tumor NCR across different multiparametric MRI scans (T1c, T2w, and FLAIR) using segmentations obtained from the RadCNN pipeline. CoLIAGE computes higher order statistics from the gradient orientation changes computed across X, Y , and Z directions in localized regions of interest. These features have been shown to be successful in tumor characterization for a variety of applications in the brain,²⁶ lung, and breast cancers, and detailed algorithmic implementation of the features in 2-D and 3-D can be found in Ref. 27.

The 3-D CoLIAGE is associated with two dominant directions, θ and ϕ , computed using singular value decomposition of gradient magnitude matrix in a locally defined window. Two separate $\mathcal{N} \times \mathcal{N}$ co-occurrence matrices, $\mathcal{M}^\theta$, and $\mathcal{M}^\phi$, corresponding to the two principal orientations, are computed. We then individually compute 13 Haralick statistics as $[S_{\phi_b}]$,

$b \in [1, 13]$ from $\mathcal{M}^\theta$ and $\mathcal{M}^\phi$, for every voxel $c \in \{C_{ET}, C_{ED}, C_{NEC}\}$, as shown in Ref. 28. For every b , first-order statistics (i.e., mean, median, standard deviation, skewness, and kurtosis) are then computed by aggregating $[S_{\theta_b}, S_{\phi_b}]$ for every $c \in \{C_{ET}, C_{ED}, C_{NEC}\}$ yielding a feature descriptor $\mathbb{F}$ for the ET, necrotic core, and peritumoral edematous region, respectively, for T1c, T2w, and FLAIR protocols.

After extraction of CoLIAGE features from every subcompartment, we used Wilcoxon rank-sum test to identify the most distinguishing features between short-term versus long-term, short-term versus mid-term, and mid-term versus long-term categories. Multilinear regression was used to predict the survival days. The top five features, along with age, are used as independent variables for survival analysis.

2.8 Performance Measures

2.8.1 Subcompartment segmentation

To evaluate the performance of our predicted segmentations relative to the GT segmentations, we computed sensitivity, specificity, dice coefficient,²⁹ and Hausdorff distance³⁰ for each tumor subregion (NCR + NET, ET, and ED). In particular, the mean, median, and standard deviation of the training, validation, and test sets were calculated through the online evaluation platform.³¹ We similarly computed dice similarity coefficient (DSC) using sensitivity (the proportion of positives that are correctly identified) and specificity (the proportion of negatives that are correctly identified) measures as follows:

$$DSC = \frac{2|GT \cap S|}{|GT| + |S|}, \quad (1)$$

which measures the extent of the spatial overlap between the GT and predicted segmentation mask, S . DSC values range between 0 (for no overlap) and 1 (perfect overlap). Similarly, the Hausdorff distance, a metric of how far two topological objects are from each other, was used to provide an additional measure of the maximum spatial distance between the GT and S . Hausdorff distance (d_H) between S and GT is given by

$$d_H(GT, S) = \max(\sup_{a \in S} \inf_{b \in GT} \|a - b\|, \sup_{b \in GT} \inf_{a \in S} \|a - b\|), \quad (2)$$

where $\|a - b\|$ is the Euclidean distance between two points, a and b belonging to GT and S , respectively. Note that a smaller d_H indicates better segmentation as the two objects are topologically closer to one another.

2.8.2 Survival analysis

The evaluation measures used in the BraTS challenge to assess the performance of the survival analysis model included accuracy (Acc), mean-squared error (MSE), and Spearman correlation coefficient (SpearmanR).³² Since the survival analysis is a prediction task (i.e., to predict the survival in terms of number of days), we calculate the MSE between the predicted survival, $\hat{y}$, and the observed GT values, Y , as follows:

$$MSE = \frac{1}{n} \sum_{i=1}^n (Y_i - \hat{y}_i)^2. \quad (3)$$

Here, a smaller MSE indicates better prediction, with 0 indicating a perfect match between the predicted survival and GT

across all samples in the dataset. Finally, to measure the statistical dependence between the predicted and GT pairs of observations, we compute SpearmanR coefficient, r_s , using

$$r_s = \frac{\text{cov}(rg_X, rg_Y)}{\sigma_{rg_X} \sigma_{rg_Y}}, \quad (4)$$

where $\text{cov}(rg_X, rg_Y)$ describes the covariance between the predicted and GT variables, and $\sigma_{rg_X}, \sigma_{rg_Y}$ are the standard deviations of the variables. Note the value $r_s = 1$ is a perfect positive correlation and $r_s = -1$ is a perfect negative correlation between the variables.

Since this is a three-class classification problem, where the predicted output is the number of days, it is therefore important to get an estimate of the deviation of the predicted output from the GT survival time. In particular, the mean, median, and standard deviation of the MSE were presented for the training, validation, and test sets, respectively.

3 Experiments and Results

3.1 Data and Preprocessing

The training dataset provided by BraTS 2017 challenge^{1,33} consists of 210 multimodal MRI (T1, T1c, T2w, FLAIR) scans of patients with HGGs and 75 patients with low-grade gliomas (LGGs). The images were skull-stripped, coregistered to a common space, and resampled to a 1-mm³ voxel resolution, with the final dimensions of each volume being 240 × 240 × 155 voxels. Each volume was normalized by subtracting the mean and dividing by the standard deviation of the intensities. The studies were also affine-aligned to the same space to ensure consistency between the data. The GT provided for each voxel consists of one of the following four subregions: NCR + NET—green, ED—yellow, and ET—blue. The GT label provided for each voxel was then used to train our model during the training portion of the pipeline. We evaluate the scans included in the test set using the same protocol, with the output for each voxel being one of the four labels described above. However, we bin the four labels into the following three regions: whole tumor, WT (NCR + NET + ED + ET), tumor core, TC (NCR + NET + ET), and ET only, to report the metrics (mean, median, and standard deviation) for the segmentation to be consistent with the BraTS challenge.

Similarly, the images within the validation and testing dataset, consisting of a total of 46 and 147 cases, respectively, were also preprocessed using the same pipeline as used for the training cohort. The GT was not included for evaluation. Instead, the evaluation was performed on the online CBICA portal provided by the University of Pennsylvania³¹ with the official results being obtained by combining the predicted segmentation into three labels: WT (NCR + NET + ED + ET), TC (NCR + NET + ET), and just the ET.

The survival criteria, using OS in months from date of baseline diagnosis, for the BraTS 2017 challenge, are as follows: (a) short-term survival with OS < 10 months ($N = 65$), (b) mid-term survival with 10 months < OS < 15 months ($N = 42$), and (c) long-term survival with OS > 15 months ($N = 56$). The validation and testing cohorts for the survival analysis had $N = 33$ and $N = 95$ studies, respectively, for which the OS information was not provided.

3.2 Results

3.2.1 Segmentation

To evaluate the impact of using textural radiomic maps in conjunction with intensity-only scans as inputs to a CNN model (RadCNN versus DeepMedic), we first trained, validated, and tested the DeepMedic CNN on a subset of the HGG and LGG cases from the BraTS 2017 training, and validation dataset, using the T1c, T2w, and FLAIR protocols. We then evaluated RadCNN on the same data (i.e., using the same cases and parameters for training, validation, and testing). We used Wilcoxon-rank sum test to compute the statistical significance between DeepMedic and RadCNN. Table 1 shows the results gathered from the online evaluation platform³¹ for both models. We also compared the DSC values of the RadCNN model with those obtained from the other models in the BraTS 2016 challenge using the corresponding 2016 dataset. The top features identified by the feature selection algorithm included Haralick entropy, energy, inverse difference moment, and correlation co-occurrence statistics. It may be observed from Table 2 that RadCNN resulted in an increase in the performance across most of the top-performing measures, except the method presented in Dera et al.³⁴ Similarly, Table 3 details the performance of our model on the independent BraTS 2017 testing dataset.

Table 1 Performance of RadCNN model (trained on intensities + radiomics features as inputs) compared with DeepMedic (trained on intensities alone). Each three-tuple represents the respective average scores for (ET, WT, and TC) across the validation cohort ($N = 46$).

	DSC	Sensitivity	Specificity	Hausdorff
DeepMedic	(0.70, 0.88, 0.73)	(0.75, 0.89, 0.71)	(0.998, 0.994, 0.997)	(5.69, 11.99, 11.78)
RadCNN	(0.71, 0.89, 0.73)	(0.75, 0.89, 0.70)	(0.998, 0.995, 0.998)	(5.24, 6.53, 10.07)

Table 2 DSC scores of top performing methods as reported in the BraTS 2016 challenge and RadCNN using the BraTS '16 training dataset (highlighted in bold).

	WT	TC	ET
Chang	0.87	0.81	0.72
Dera	0.91	0.91	0.84
Krishnamurthi	0.84	0.71	0.81
Randhawa	0.87	0.75	0.81
Song	0.86	0.70	0.73
Vilaplana	0.89	0.76	0.37
Zeng	0.85	0.82	0.80
Zhao	0.87	0.82	0.76
DeepMedic	0.90	0.75	0.72
RadCNN	0.90	0.82	0.80

Table 3 Performance of RadCNN on the BraTS '17 testing set. Each three-tuple represents the respective average scores for (ET, WT, and TC) across the test cohort ($N = 147$).

	DSC	Hausdorff
Mean	(0.68, 0.86, 0.74)	(45.00, 7.03, 33.08)
Standard deviation	(0.30, 0.13, 0.29)	(116.06, 10.69, 93.95)

Compared to the intensity-based DeepMedic model, the proposed pipeline showed an improvement in both the enhancing and whole tumor DSC and obtained comparable results in segmenting the TC. Although the specificity of the TC does

decrease slightly when texture features are incorporated, the Hausdorff distance between the prediction and the GT decreases significantly (about 50% for the whole tumor and 14% for the TC). This suggests that the addition of preselected textural radiomic maps may result in an overall improvement in the predicted segmentation. Figure 2 shows examples from three different patient studies, where there was an improvement in subcompartment segmentations using RadCNN compared to using just DeepMedic. While the dice, sensitivity, and specific measures across Deepmedic and RadCNN were not statistically significantly different, statistically significant differences were observed for the Hausdorff distance measure across both the training (0.046) and validation (0.04) sets.

Figure 3 shows a case taken from the BraTS 2017 training dataset³³ that benefits from using connected component

Fig. 2 Examples of segmentation results. From left to right: GT, RadCNN segmentation prediction, and an intensity-only DeepMedic CNN segmentation prediction. Green, blue, and yellow labels denote the NCR + NET, ET, and ED, respectively.

Fig. 3 Comparison between (a) GT, (b) RadCNN (no postprocessing), and (c) RadCNN (connected component analysis). The connected component analysis removed the ED segmentation (shown inside red circle) on the right-half of the brain (radiological view), making the predicted segmentation more accurate compared to RadCNN alone (i.e., 81.5% accuracy versus 78.6% for the illustrated case).

Fig. 4 Comparison between (a) GT, (b) RadCNN (no postprocessing), and (c) RadCNN (MRF). The postprocessed MRF prediction map filled some holes within the ED segmentation while also removing some incorrect NCR + NET segmentations along the boundary between ED and ET.

analysis. This technique was able to remove the far disconnected regions from the main tumor, which showed an improvement to the ED segmentation. Overall, this resulted in an improvement of about 3% to the WT DSC score from 78.6% to 81.5%.

Similarly, Fig. 4 shows a case taken from the BraTS 2017 training dataset³³ that shows the difference in segmentation compared to GT when applying an MRF to the prediction map $\mathcal{P}$. This technique was able to close some gaps within the lower-left ED segmented region as well as correcting some incorrectly predicted NCR + NET segmentations along the boundary voxels between ED and ET, resulting in an improvement of about 1.5% to the WT DSC score from 81.2% to 82.8%.

To compare the effects of selecting texture features on each individual label, the percentage of under- and oversegmentation of each label on a per-patient basis was calculated. With the addition of texture features, the undersegmentation percentages of each label tend to decrease overall. On average, for the cross-validation cases, this error rate decreased by 1% (20% versus 19%) for the ED label and shows comparable results with respect to NCR + NET and ET labels.

Since U-Net is one of the standard segmentation architectures used recently in medical image analysis, both DeepMedic and RadCNN were compared against a U-Net trained model

modified to work with 3-D MRI scans. To adequately and accurately compare RadCNN with U-Net, we created two different models: (1) using raw intensities alone and (2) using both intensity and texture channels (similar to RadCNN). We utilized the same texture maps that were used within RadCNN. The results are summarized in Table 4. Our results seem to suggest that the performance of RadCNN, with regard to both the DSC and Hausdorff distance, was better than U-Net trained on (a) intensities alone and (b) intensities + texture, across all three compartments, ET, WT, and TC, respectively.

3.2.2 Survival analysis

The top features identified for the survival prediction were the peritumoral CoLIAGE descriptors from the FLAIR protocol. Qualitative maps depicting the expression levels of the best feature (CoLIAGE entropy) on FLAIR sequence scans, across a representative short-, mid- and long-term survivor is shown in Fig. 5. As may be observed, CoLIAGE features exhibited differential expression in the peritumoral ED region. However, we did observe significant differences in CoLIAGE expressions in the intratumoral or necrotic zones. Training, validation, and testing performance metrics are shown in Table 5.

Table 4 Performance of RadCNN compared to the conventional 23 layer U-Net architecture. Each entry represents the mean score and standard deviation for ET, WT, and TC across the validation cohort ($N = 46$).

Architecture	DSC (mean/Std)			Hausdorff distance (mean/Std)		
	ET	WT	TC	ET	WT	TC
RadCNN	0.71/0.30	0.89/0.077	0.73/0.29	5.42/10.08	6.53/8.75	10.07/14.11
U-net (intensities alone)	0.67/0.33	0.87/0.088	0.68/0.31	8.42/18.19	15.25/21.29	16.96/26.71
U-net (intensities + texture)	0.68/0.32	0.88/0.085	0.72/0.27	8.97/18.82	10.55/16.42	14.07/24.45

Fig. 5 CoLIAGE expression maps: (a)–(c) a representative long-term, mid-term and short-term survivor along with the corresponding GT expert annotations. Corresponding CoLIAGE entropy maps are shown in (d)–(f), respectively. Even though the features have similar expressions in the ET and NEC regions, there is a clear difference observed in the peritumoral zone, which was leveraged to build the prognostic model.

Table 5 Performance metrics for BraTS 2017 survival analysis task.

	Accuracy	MeanSE	MedianSE	StdSE	SpearmanR
Training	0.57	214880.667	26156.426	555254.501	0.441
Validation	0.63	204362.859	20350.2	524463.072	0.457
Test	0.453	250644.056	35601.592	768370.877	0.38

4 Discussion and Conclusion

In this paper, we presented RadCNN approach to improve brain subcompartment segmentation, by providing optimized radiomic representations of the multiparametric MRI scans as inputs to a 3-D CNN classifier. We also used an optimized set of radiomic measurements from the tumor subcompartments to build a model, prognostic of survival, which was evaluated on the BraTS 2017 test set. Our results suggest that (1) radiomic

features, owing to their ability to provide complementary phenotypic information,³⁵ in conjunction with 3-D CNN, can augment lesion segmentation performance; and (2) peritumoral radiomic descriptors are more prognostic of overall patient survival than the features extracted from the ET or necrotic zones. Radiomic texture features have been previously shown to possibly capture lesion heterogeneity.³⁵ We have previously shown that peritumoral features may be predictive of OS in HGGs. For

Fig. 6 Illustration of possible causes of error: (a) a case with partial volume effect (yellow specks inside the red circle), annotated as ED in the GT provided, is depicted. These areas were missed by RadCNN. Similarly, the regions circled in red in (b) are characterized as regions of gliomatosis. However, these regions were classified as ED by RadCNN.

example, entropy-based descriptors might reflect the underlying cell population or mitotic processes, as observed on histopathology.³⁶

RadCNN method was found to improve segmentation performance across both cohorts from BraTS 2016 and 2017 challenges, except when compared with segmentation results from Dera et al.³⁴ on BraTS 2016 cohort. Dera et al. employed non-negative matrix factorization level set segmentation algorithm, which is known to be robust to intensity inhomogeneity extent in MRI scans and hence may have produced improved segmentation results than RadCNN.

As shown in Fig. 6, we identified a few possible sources of error for RadCNN. Figure 6 depicts a case with partial volume effect (specks inside the red region), annotated as ED in the GT provided. These areas were not successfully identified by RadCNN as ED. Similarly, the regions circled in red in Fig. 6 are characterized as regions of gliomatosis. Interestingly, these regions were classified as ED by RadCNN, therefore resulting in lower DSC scores.

One possible reason for the difference in accuracy between the training and test sets could be the large variation in clinical parameters across the multi-institutional training and test cohorts. However, in the absence of any clinical- or acquisition-related details, we cannot make any substantial claims regarding the exact reason for the difference in accuracy across training and testing. Future work comprises further optimizing the input channels by incorporating additional low-level hand-crafted features (such as shape characteristics, distance between subcompartment centroids) to help improve segmentation predictions. These features will be evaluated and compared to our current results to create a more optimized feature set. Furthermore, we intend to incorporate MRF into the CNN framework and also combine segmentations from different models to create more robust segmentation predictions. For the survival task, the only clinical feature used in conjunction with the peritumoral radiomic ones was the age of the patient at the time of initial diagnosis. We believe that the inclusion of additional clinical parameters, such as Karnofsky performance score,³⁷ can potentially result in more accurate prognostic models.

Disclosures

Dr. Madabhushi is an equity holder in Elucid Bioimaging and in Inspirata, Inc. He is also a scientific advisory consultant for Inspirata, Inc. and also sits on its scientific advisory board. Additionally, his technology has been licensed to Elucid Bioimaging and Inspirata, Inc. He is also involved in a NIH U24 grant with PathCore, Inc. His work is also sponsored by Philips.

Acknowledgments

Research reported in this publication was supported by the National Cancer Institute of the National Institutes of Health under award numbers 1U24CA199374-01, R01CA202752-01A1, R01CA208236-01A1, R01 CA216579-01A1 R01 CA220581-01A1; National Center for Research Resources under award number 1 C06 RR12463-01; Merit Review Award VA IBX004121A from the United States (U.S.); Department of Veterans Affairs Biomedical Laboratory Research and Development Service the DOD Prostate Cancer Idea Development Award (W81XWH-15-1-0558); the DOD Lung Cancer Investigator-Initiated Translational Research Award (W81XWH-18-1-0440); the DOD Peer Reviewed Cancer Research Program (W81XWH-16-1-0329); the Ohio Third Frontier Technology Validation Fund; the Wallace H. Coulter Foundation Program in the Department of Biomedical Engineering and the Clinical and Translational Science Award Program (CTSA) at Case Western Reserve University. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health or U.S. Department of Veterans Affairs or the United States Government.

References

1. B. H. Menze et al., "The multimodal brain tumor image segmentation benchmark (BraTS)," *IEEE Trans. Med. Imaging* **34**(10), 1993–2024 (2015).
2. G. P. Mazzara et al., "Brain tumor target volume determination for radiation treatment planning through automated MRI segmentation," *Int. J. Radiat. Oncol. Biol. Phys.* **59**(1), 300–312 (2004).
3. T. Yanagihara et al., "A simple automated method for detecting recurrence in high-grade glioma," *Am. J. Neuroradiol.* **37**(11), 2019–2025 (2016).
4. K. Kamnitsas et al., "Efficient multi-scale 3D CNN with fully connected CRF for accurate brain lesion segmentation," *Med. Image Anal.* **36**, 61–78 (2017).
5. T. Lun and W. Hsu, "Brain tumor segmentation using deep convolutional neural network," in *Proc. BraTS-MICCAI* (2016).
6. R. Meier et al., "CRF-based brain tumor segmentation: alleviating the shrinking bias," in *Int. Workshop on Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries*, Springer, pp. 100–107 (2016).
7. S. Luan et al., "Gabor convolutional networks," *IEEE Trans. Image Process.* **27**(9), 4357–4366 (2018).
8. V. John, A. Boyali, and S. Mita, "Gabor filter and Gershgorin disk-based convolutional filter constraining for image classification," *Int. J. Mach. Learn. Comput.* **7**(4), 55–60 (2017).
9. Y. Chen et al., "Hyperspectral images classification with Gabor filtering and convolutional neural network," *IEEE Geosci. Remote Sens. Lett.* **14**, 2355–2359 (2017).
10. Y. J. Ryu et al., "Glioma: application of whole-tumor texture analysis of diffusion-weighted imaging for the evaluation of tumor heterogeneity," *PLoS One* **9**(9), e108335 (2014).
11. P. Prasanna, P. Tiwari, and A. Madabhushi, "Co-occurrence of local anisotropic gradient orientations (CoLIAGe): a new radiomics descriptor," *Sci. Rep.* **6**, 37241 (2016).
12. M. Goetz et al., "Extremely randomized trees based brain tumor segmentation," in *Proc. BraTS Challenge-MICCAI*, pp. 006–011 (2014).

13. R. R. Rawat et al., "Correlating nuclear morphometric patterns with estrogen receptor status in breast cancer pathologic specimens," *NPJ Breast Cancer* **4**(1), 32 (2018).
14. H. Wang et al., "Mitosis detection in breast cancer pathology images by combining handcrafted and convolutional neural network features," *J. Med. Imaging* **1**(3), 034003 (2014).
15. R. Stupp et al., "Concomitant and adjuvant temozolomide (TMZ) and radiotherapy (RT) for newly diagnosed glioblastoma multiforme (GBM). Conclusive results of a randomized phase III trial by the EORTC Brain & RT Groups and NCIC Clinical Trials Group," *J. Clin. Oncol.* **22**(14_suppl), 2–2 (2004).
16. D. Krex et al., "Long-term survival with glioblastoma multiforme," *Brain* **130**(10), 2596–2606 (2007).
17. W. A. Osta et al., "EPCAM is overexpressed in breast cancer and is a potential target for breast cancer gene therapy," *Cancer Res.* **64**(16), 5818–5824 (2004).
18. N. M. Braman et al., "Intratumoral and peritumoral radiomics for the pretreatment prediction of pathological complete response to neoadjuvant chemotherapy based on breast DCE-MRI," *Breast Cancer Res.* **19**(1), 57 (2017).
19. A. K. Jain and F. Farrokhnia, "Unsupervised texture segmentation using Gabor filters," *Pattern Recognit.* **24**(12), 1167–1186 (1991).
20. K. I. Laws, "Texture energy measures," in *Proc. Image understanding Workshop*, pp. 47–51 (1979).
21. R. M. Haralick et al., "Textural features for image classification," *IEEE Trans. Syst. Man Cybernet.* **3**(6), 610–621 (1973).
22. H. Peng, F. Long, and C. Ding, "Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy," *IEEE Trans. Pattern Anal. Mach. Intell.* **27**(8), 1226–1238 (2005).
23. D. Kingma and J. Ba, "ADAM: a method for stochastic optimization," arXiv:1412.6980 (2014).
24. L. Vincent, "Morphological grayscale reconstruction in image analysis: applications and efficient algorithms," *IEEE Trans. Image Process.* **2**(2), 176–201 (1993).
25. S. Geman and D. Geman, "Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images," *IEEE Trans. Pattern Anal. Mach. Intell.* **PAMI-6**(6), 721–741 (1984).
26. P. Prasanna et al., "Disorder in pixel-level edge directions on T1WI is associated with the degree of radiation necrosis in primary and metastatic brain tumors: preliminary findings," *Am. J. Neuroradiol.* **40**(3), 412–417 (2019).
27. P. Prasanna, P. Tiwari, and A. Madabhushi, "Co-occurrence of local anisotropic gradient orientations (collage): distinguishing tumor confounders and molecular subtypes on MRI," in *Int. Conf. Med. Image Comput. Comput.-Assist. Interv.*, Springer, pp. 73–80 (2014).
28. P. Prasanna et al., "Radiographic-deformation and textural heterogeneity (r-depth): an integrated descriptor for brain tumor prognosis," in *Int. Conf. Med. Image Comput. Comput.-Assist. Interv.*, Springer, pp. 459–467 (2017).
29. L. R. Dice, "Measures of the amount of ecologic association between species," *Ecology* **26**(3), 297–302 (1945).
30. R. T. Rockafellar and R. J.-B. Wets, *Variational Analysis*, Vol. **317**, Springer Science & Business Media, Berlin, Germany (2009).
31. Center for Biomedical Imaging and Computing Analytics, Univ. of Pennsylvania, "CBICA image processing portal," 2015, <https://ipp.cbica.upenn.edu/> (accessed April 2019).
32. M. M. Mukaka, "A guide to appropriate use of correlation coefficient in medical research," *Malawi Med. J.* **24**(3), 69–71 (2012).
33. S. Bakas et al., "Advancing the cancer genome atlas glioma MRI collections with expert segmentation labels and radiomic features," *Sci. Data* **4**, 170117 (2017).
34. D. Dera, N. Bouaynaya, and H. Fathallah-Shaykh, "Assessing the non-negative matrix factorization level set segmentation on the BraTS benchmark," in *Proc. MICCAI-BraTS Workshop 2016*, pp. 10–13 (2016).
35. P. Prasanna et al., "Radiomic features from the peritumoral brain parenchyma on treatment-naïve multi-parametric MR imaging predict long versus short-term survival in glioblastoma multiforme: preliminary findings," *Eur. Radiol.* **27**(10), 4188–4197 (2017).
36. T. Homma et al., "Correlation among pathology, genotype, and patient outcomes in glioblastoma," *J. Neuropathol. Exp. Neurol.* **65**(9), 846–854 (2006).
37. J. W. Yates et al., "Evaluation of patients with advanced cancer using the Kamofsky performance status," *Cancer* **45**(8), 2220–2224 (1980).

Prateek Prasanna is a research associate at the Center for Computational Imaging and Personalized Diagnostics, Case Western Reserve University. His research involves developing quantitative image analysis tools to address unmet clinical needs and aid in precision oncology.

Ayush Karnawat is a senior undergraduate student in the Department of Electrical Engineering and Computer Science.

Marwa Ismail is a research associate at the Brain Image Computing (BrIC) Laboratory. Her research has involved several brain-related disorders, such as nerve fiber tracking using DTI, and autism diagnosis using multiple imaging modalities. Her current research is developing image analysis techniques for characterizing brain tumors.

Anant Madabhushi is the director of CCIPD and the F. Alex Nason Professor II in the Departments of Biomedical Engineering, Pathology, Radiology, Radiation Oncology, Urology, General Medical Sciences, and Electrical Engineering and Computer Science at Case Western Reserve University. He is also a member of the Case Comprehensive Cancer Center. He has authored over 150 peer-reviewed journal publications and over 180 conferences papers and delivered over 230 invited talks and lectures both in the US and abroad.

Pallavi Tiwari is an assistant professor of biomedical engineering and the director of BrIC Laboratory at Case Western Reserve University. She is also an associate member of the Case Comprehensive Cancer Center. Her research interests lie in development of novel machine learning and pattern recognition methods for neurological disorders. She has coauthored over 50 peer-reviewed journal and conference articles on her work in personalized medicine solutions across different applications.