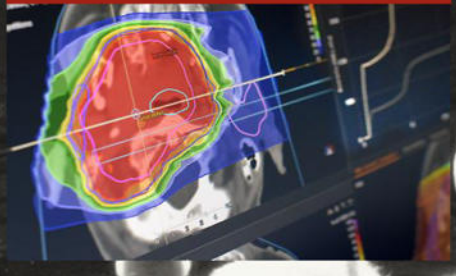
The Family

Delta^4 TPV



Delta^4 Phantom+

Delta^4 Discover

QA from prescription to final fraction

Delta^4 Family one software platform for all your QA needs.

With Quality Assurance from prescription to final fraction you can now increase your workflow efficiency and be confident that the treatment dose delivered to your patient is safe.

Delta^4
by ScandiDos

Innovative and Efficient QA

www.delta4family.com

Deep learning for segmentation of brain tumors: Impact of cross-institutional training and testing

Ehab A. AlBadawy and Ashirbani Saha
Department of Radiology, Duke University School of Medicine, Durham, NC, USA

Maciej A. Mazurowski[a)]
Department of Radiology, Duke University School of Medicine, Durham, NC, USA
Department of Electrical and Computer Engineering, Duke University, Durham, NC, USA
Duke University Medical Physics Program, Durham, NC, USA

(Received 10 April 2017; revised 13 December 2017; accepted for publication 14 December 2017; published 8 February 2018)

Background and purpose: Convolutional neural networks (CNNs) are commonly used for segmentation of brain tumors. In this work, we assess the effect of cross-institutional training on the performance of CNNs.
Methods: We selected 44 glioblastoma (GBM) patients from two institutions in The Cancer Imaging Archive dataset. The images were manually annotated by outlining each tumor component to form ground truth. To automatically segment the tumors in each patient, we trained three CNNs: (a) one using data for patients from the same institution as the test data, (b) one using data for the patients from the other institution and (c) one using data for the patients from both of the institutions. The performance of the trained models was evaluated using Dice similarity coefficients as well as Average Hausdorff Distance between the ground truth and automatic segmentations. The 10-fold cross-validation scheme was used to compare the performance of different approaches.
Results: Performance of the model significantly decreased ($P < 0.0001$) when it was trained on data from a different institution (dice coefficients: 0.68 ± 0.19 and 0.59 ± 0.19) as compared to training with data from the same institution (dice coefficients: 0.72 ± 0.17 and 0.76 ± 0.12). This trend persisted for segmentation of the entire tumor as well as its individual components.
Conclusions: There is a very strong effect of selecting data for training on performance of CNNs in a multi-institutional setting. Determination of the reasons behind this effect requires additional comprehensive investigation. [https://doi.org/10.1002/mp.12752]

Key words: brain tumor segmentation, CNN, glioblastoma, impact of cross-institutional training, magnetic resonance imaging

1. INTRODUCTION

Automatic segmentation of brain tumors from magnetic resonance imaging (MRI) data is a research topic of great interest over the last several years.[1–6] Automatic brain tumor segmentation using deep learning methodologies have garnered attention from researchers. A recent survey of deep learning-based brain tumor segmentation can be found in Ref. [7]. A general improvement in the segmentation of brain tumors has been reported in different studies[5,6,8–14] due to the use of convolutional neural networks (CNNs) which is a type of deep learning algorithm. The automatic algorithms used in these studies are applicable for patients with glioblastoma and low-grade gliomas. State-of-the-art performance has been achieved by CNNs using multimodal imaging data for segmentation. By employing two-dimensional CNNs in the publicly available Multimodal Brain Tumor Image Segmentation Benchmark (BRATS) dataset,[15] the methods[6,11,14] have achieved improved brain tumor segmentation in the recent years. Also, three-dimensional CNNs[12,13] achieved competitive segmentation performance in this dataset. BRATS dataset is a multi-institutional dataset having the data collected from four institutions[15] and has been significantly modified over several years (https://sites.google.com/site/braintumorsegmentation/).

Recent trend of sharing multi-institutional datasets facilitates research in many different aspects including repeatability of experiments and an opportunity to work with data having more variability in terms of patient population, imaging equipment, and contrast protocols. At the same time, a question arises: how does this interinstitutional variability affect the performance of the algorithms trained with this heterogeneous data? Different institutions may differ in terms of their imaging equipment, image acquisition parameters, and contrast injection protocols which might notably affect imaging characteristics making them systematically diverse between different institutions. This in turn can cause a CNN model trained with data from one institution perform poorly on data from another institution. In this study, we investigate this issue in the context of CNNs for segmenting tumors from glioblastoma (GBM) patients. Specifically, we used MRI data of glioblastoma (GBM) patients for two institutions present in The Cancer Imaging Archive.

The issue of potential decrease in segmentation performance due to interinstitutional differences is important

because of the following reasons. Segmentation of lesions is one of the preliminary steps in the automatic extraction of meaningful features which are finding increasing importance in assessing the extent of disease and prognosis of outcomes.[16–18] Thus, automatic tumor segmentation will greatly facilitate truly automated brain radiomics[19] or radiogenomics (i.e., finding associations between imaging and genomics).[20,21] If different data sources affect tumor segmentation, they will also affect the extracted features potentially impacting performance of predictive models. To properly account for it, one must first determine whether such effect occurs and measure its magnitude. This study is the first step in this direction. Previously, we have investigated the effect of cross-institutional training in the context of fully automated feature extraction from GBM[18] using random forest classifiers. However, at this juncture when deep learning-based systems are showing increasing effectiveness over classical machine learning techniques, we found the assessment of cross-institutional training performance with CNN-based segmentation to be a very relevant one. The question investigated in this work has not been studied in the context of deep neural networks for the processing of medical images and in terms of interinstitutional differences present in the data source.

2. MATERIALS AND METHODS

2.A. Dataset and ground truth for segmentation

We obtained data for 68 (who had both preoperative MRI data and survival data available) patients from two institutions from The Cancer Imaging Archive GBM dataset (TCIA, http://cancergenome.nih.gov/). We chose 22 GBM patients from MD Anderson Cancer Center and 46 GBM patients from Henry Ford Hospital for this study. Then, we randomly selected 22 cases from Henry Ford Hospital to have an equal number of cases from the two institutions which was important given our experimental design described below. We selected 22 patients from each of the institutions since this was the largest number of patients with usable data available, such that an equal number of patients from two institutions could be obtained for our analysis. Therefore, there were 44 patients available for our experiments. For each of these patients, the following imaging data were available: T1-weighted precontrast, T1-weighted postcontrast, and fluid-attenuated inversion recovery (FLAIR) sequences. The sequences were registered (to the sequence with the least number of slices) using the FLIRT registration tool of FSL (FMRIB Software Library) software (http://fsl.fmrib.ox.ac.uk/fsl/fslwiki/FLIRT).[22] A manual segmentation of the tumors was generated using the following steps: (a) an in-house graphical user interface was used by a researcher in our laboratory to manually delineate the tumor, (b) these annotations were approved or corrected as necessary by a neuroradiology fellow to generate the final ground truth. The final ground truth consisted of five classes: normal appearing brain (Class 0), enhancing region (Class 2), necrotic region (Class 3), T1 abnormality (hypointensity region on T1) excluding enhancing and necrotic regions (Class 4), and FLAIR abnormality (hyperintensity region on FLAIR) excluding the already mentioned classes (Class 5). The union of regions represented by Classes 2–5 is used to generate the ground truth for the entire tumor region (Class-1).

2.B. Convolutional neural network

Convolutional neural network generally consists of a combination of different layers. Typically, the first layers are convolutional layers which contain linear filters with learnable weights, followed by nonlinear activation. For every filter, there is an output which called a feature map. Pooling layer is used in between the other layers of the network to reduce the spatial size for the feature maps for computation efficiency and for controlling overfitting. At the end, a fully connected layer is used with a dropout layer that turns off some neurons during training to reduce overfitting. For our experiments, we used the CNN shown in Fig 1. We used a similar CNN model to the one mentioned in Ref. [8] with a few changes as this model was placed among the top performers in the BRATS challenge 2015.[15] The input for this model is a three-dimensional image, called patch. First, we had input patch of size $33 \times 33 \times 3$ as shown in Fig. 2, where 33×33 are patch dimensions, with the three sequences (precontrast, postcontrast, and FLAIR). Four convolutional layers were used for feature extraction from the patches by convolving a set of weights, organized as kernels. We used Rectified Linear Unit (ReLU)[23] as the activation function, which imposes a constant 0 in negative input and linear output with slope 1 otherwise. Two max-pooling layers (indicated by blue arrows in Fig. 1) were employed with some overlapping of the receptive fields, to keep important details for segmentation. At the end of the network, there were two fully connected layers with dropout $P = 0.5$ as regularization, in order to reduce overfitting and one SoftMax layer with five outputs represents target labels. For training, a constant learning rate 0.0003 was used for 25 epochs. The details of patch generation and selection are discussed next.

A patch is a part of the image with a label corresponding to a class of its center voxel. The three channels in the patch were formed using the corresponding regions of precontrast, postcontrast, and FLAIR sequence, respectively. We used a $33 \times 33 \times 3$ patch size. The channel size of 33×33 has been used in earlier approaches.[8,9] An example patch with the three channels is shown in Fig 2. In order to select patches for training, we divided the patches into three categories, as shown in Fig. 3: (a) patches belonging to the tumor, (b) healthy patches immediately surrounding the tumor, and (c) randomly selected healthy patches from the rest of the brain.

For each patient included in the training set, we selected every tumor patch and healthy surrounding patch available. The rest of the patches were selected randomly from the remaining healthy brain to reach the total number of 3 million patches from all cases in each institution for the training of the CNN. Approximately, in the set of patches extracted, 84% (MD Anderson Cancer Center) and 75% (Henry Ford

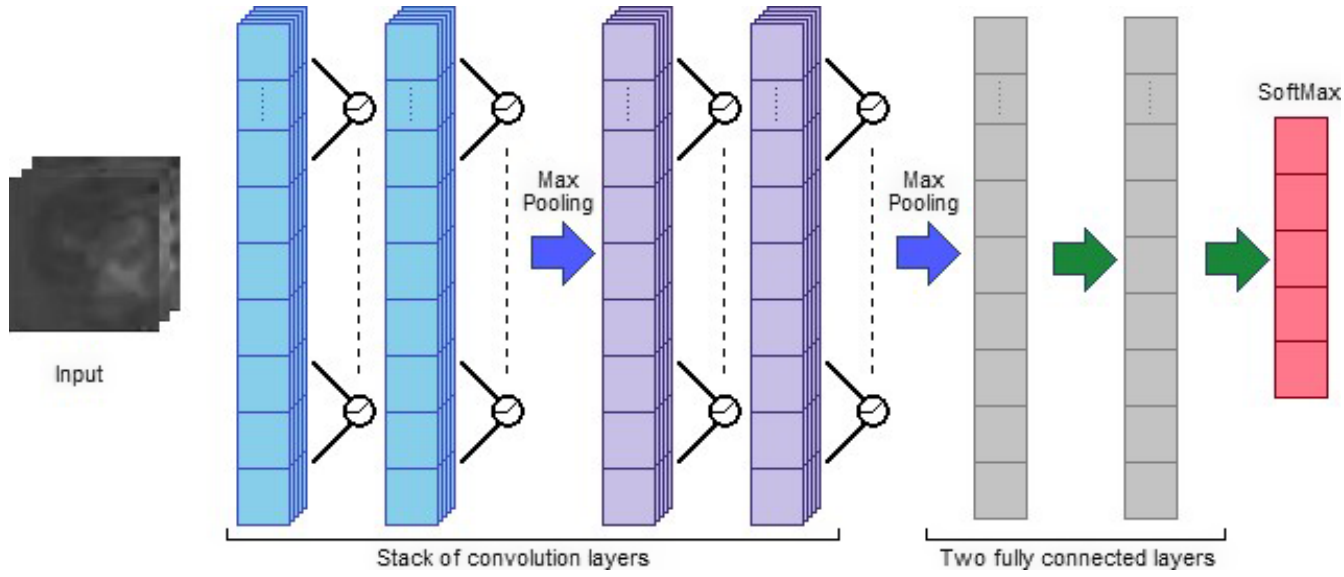


FIG. 1. A visualization of the convolutional neural network used in our experiments. Four convolution layers followed by two fully connected layers were used. The arrows denote the different operations. [Color figure can be viewed at wileyonlinelibrary.com]

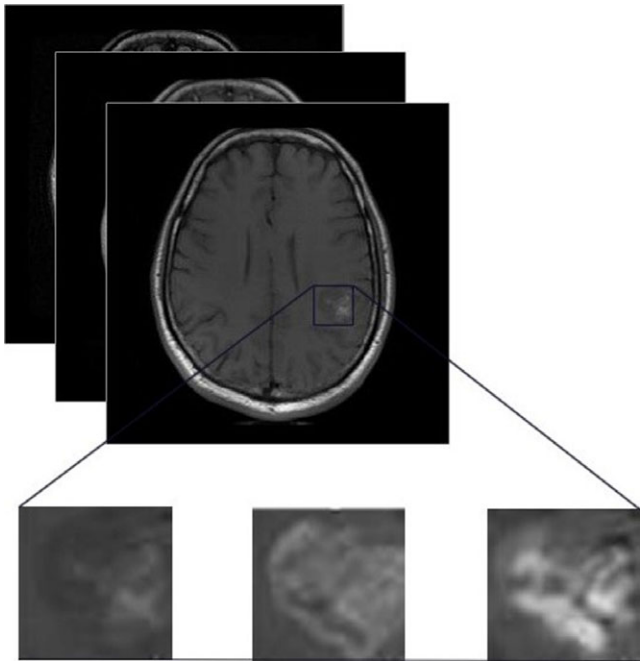


FIG. 2. Demonstrating magnified regions in precontrast, postcontrast, and FLAIR sequences, respectively, used to form a three-dimensional input patch representing a tumor. [Color figure can be viewed at wileyonlinelibrary.com]

Hospital) of patches belonged to healthy brain voxels. The rest of the patches belonged to different tumor categories. We trained multiclass CNNs, to segment brain images into the five classes described above (0, 2, 3, 4, and 5).

2.C. Design of experiments

Our goal was to evaluate three scenarios that reflect a common practice in research and evaluation of machine learning model:

FIG. 3. Patch extraction map. (a) tumor voxels in white, (b) healthy surrounding voxels to the tumor in gray, and (c) randomly selected healthy voxels in black within the healthy brain as the center of healthy patches.

(a) Data used for training of a model comes from the same institution as the data used for testing of the model
(b) Data used for training of a model comes from different institution than the data used for testing of the model
(c) Data used for training of a model that comes from the same institution as the data used for testing of the model is enriched by additional data coming from a different institution resulting in an increased size of the training set.

Using 10-fold cross-validation, three scenarios were simulated in the following way: First, we numbered the patients serially for each institution. Within a fold, the serial numbers for training and test set were obtained. For automatically segmenting the patients in the test set, we developed three CNNs

in the following manner. (a) The first CNN was trained on the training images from the same institution using the serial numbers in the training set. (b) The second CNN was trained on the patients using the training set serial numbers from the other dataset. In this way, we used the same number of patients used for training the CNN. (c) The third CNN was built using all the patients used in (a) and (b) such that patients from both institutions were used. Therefore, we used six types of train-test combinations: (a) Train on Institution 1 (I1) — Test on Institution 1 (I1), (b) Train I2 — Test I2, (c) Train I1 — Test I2, (d) Train I2 — Test I1, (e) Train (I1 + I2) — Test I1, and (f) Train (I1 + I2) — Test I2. Please note that the data for a patient whose images were segmented by a CNN was never present in the training set for that CNN. For each patient under test, we classified all the voxels of its images into one of the five classes (normal brain and specific tumor classes).

For additional validation, we conducted an experiment where a trained model generation used the maximum number (N = 24) of randomly selected cases from the institution (I2) and evaluated it on an equal number of unseen cases in I2 and I1 (22 for each).

2.D. Evaluation of segmentation

For each patient and for each tumor class (classes 1–5), we calculated three Dice coefficients that correspond to the three training-testing scenarios described above. Dice coefficient is commonly for evaluating the accuracy of segmentation.[24] Each of these Dice coefficients was calculated between the manual segmentation and (a) the segmentation obtained by training-testing within the same institution, (b) the segmentation obtained by training-testing across institutions, (c) the segmentation obtained by training-testing with both institutions. The mean value of the Dice coefficients for each institution was calculated for each type of segmentation. In order to determine whether there is a significant difference in the segmentation performance of each class between different training-testing scenarios, we performed a paired *t*-test between the Dice coefficients in the different scenarios. The same experiment was evaluated with Average Hausdorff Distance (AHD), and a combination of Sensitivity and Positive Predictive Value (PPV) using open-source Evaluate Segmentation (https://github.com/Visceral-Project/EvaluateSegmentation) code. In medical images, it is very likely to have noise and outliers and it is well known that Hausdorff Distance (HD) is sensitive to these outlines, so different modifications[25] were proposed to overcome such a problem. In our experiments, we decided to use AHD as one of these modifications. For a given two sets of points A and B, AHD is calculated as:

$$AHD = max(d_{AHD}(A, B), d_{AHD}(B, A))$$

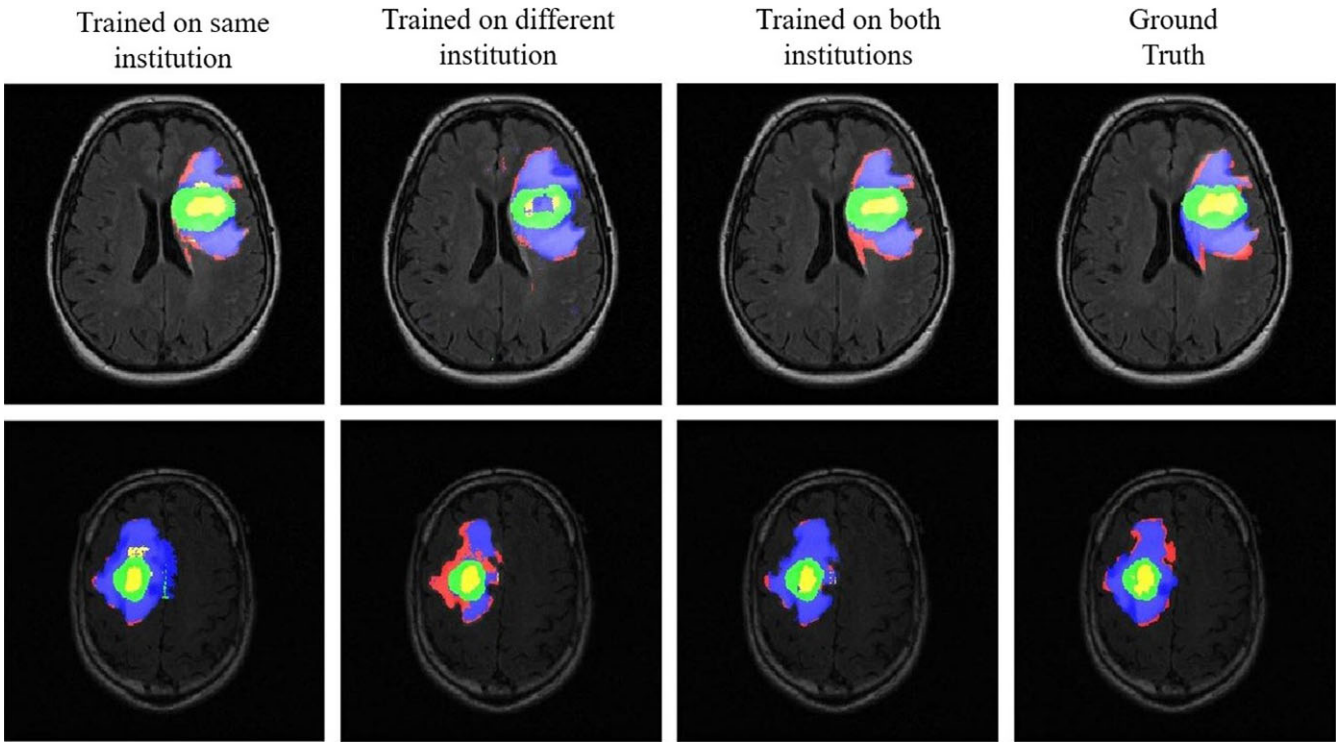


FIG. 4. Segmentation results on two selected cases from both institutions. The first three columns represent segmentation using models trained on same, different, and both institutions. The last column is the ground truth. Class 2 (enhancing region) is segmented in green, Class 3 (necrotic region) in yellow, Class 4 in red (T1 abnormality — hypointensity region on T1, excluding enhancing and necrotic regions), and Class 5 (FLAIR abnormality excluding classes 2–4) in blue. [Color figure can be viewed at wileyonlinelibrary.com]

where $d_{AHD}(A,B)$ is the directed AHD that is given by:

$$d_{AHD}(A,B) = \frac{1}{N}\sum_{a\in A}\min_{b\in B}\|a-b\|$$

However, HD is calculated as:

$$HD = max(d_{HD}(A,B), d_{HD}(B,A))$$

where $d_{HD}(A,B)$ is the directed Hausdorff distance that is given by:

$$d_{HD}(A,B) = \max_{a\in B}\min_{b\in B}\|a-b\|$$

3. RESULTS

The automatic segmentation results in Fig. 4 show two example cases: one form MD Anderson Cancer Center (first row) and one from Henry Ford Hospital (second row). Visually, we find that the segmentation results are closer to ground truth when data from the same institution, and data from both institutions were used to generate the trained models.

The average values of the Dice coefficients for each class by training-testing within the same institution, across institutions, with both institutions, are shown in Table I.

We found that the performance of the model for segmentation of the entire tumor significantly decreases ($P < 2.3 \times 10^{-6}$) when the model was trained on data from a different institution (dice coefficient = 0.68 ± 0.19 for I1 and dice coefficient = 0.59 ± 0.19 for I2) as compared to when it was trained with data from the same institution (dice coefficient = 0.72 ± 0.17 for I1 and dice coefficient = 0.76 ± 0.12 for I2). This effect persisted when analyzing the segmentation performance for the particular tumor classes with the strongest effect for Class 3 (dice coefficient = 0.50 ± 0.32 for I1 and dice coefficient = 0.51 ± 0.29 for I2 for training with the data from the same institution vs dice coefficient = 0.19 ± 0.24 for I1 and dice coefficient = 0.34 ± 0.27) for training with the data from different institutions.

Adding data from a different institution to already existing data from the institution where the model will be tested resulted in no significant difference in terms of the segmentation of the entire tumor ($P = 0.22$), and Class 4 ($P = 0.63$). Minor differences were seen for Classes 2, 3, and 5 ($P < 0.05$) for training with same vs both institutions. For training with different institutions vs both institutions, significant improvements were shown for Class 1 ($P < 1.2 \times 10^{-6}$), and Class 2, 3, 4 ($P < 1 \times 10^{-5}$).

Similar to dice coefficient, AHD shows that training the model on different institution decreased the performance (AHD = 4.98 mm for I1 and AHD = 1.36 mm for I2) compared to training on the same institution dataset (AHD = 2.58 mm for I1 and AHD = 1.0 mm for I2). Adding training dataset from both institutions improve the performance slightly. Full results using AHD are reported in Table II.

TABLE I. Average dice coefficients for different classes for different types of train-test combinations using 10-fold cross-validation.

	Class 1		Class 2		Class 3		Class 4		Class 5	
Train	Test on I1	Test on I2	Test on I1	Test on I2	Test on I1	Test on I2	Test on I1	Test on I2	Test on I1	Test on I2
Same institution	0.72 ± 0.17	0.76 ± 0.12	0.60 ± 0.23	0.68 ± 0.18	0.50 ± 0.32	0.51 ± 0.29	0.26 ± 0.09	0.20 ± 0.08	0.39 ± 0.22	0.57 ± 0.16
Different institutions	0.68 ± 0.19	0.59 ± 0.19	0.49 ± 0.29	0.51 ± 0.19	0.19 ± 0.24	0.34 ± 0.27	0.18 ± 0.09	0.16 ± 0.12	0.37 ± 0.20	0.29 ± 0.19
Both institutions	0.70 ± 0.18	0.75 ± 0.13	0.55 ± 0.28	0.68 ± 0.18	0.30 ± 0.28	0.54 ± 0.20	0.20 ± 0.12	0.21 ± 0.08	0.39 ± 0.19	0.56 ± 0.18

A combination of Sensitivity and Positive Predictive Value (PPV) has been used as another metric of evaluation. We found that when training on different institution either sensitivity or PPV values are less than 0.65. However, for training on the same or both institutions, the sensitivity and PPV values were always greater than 0.65. Specifically, low sensitivity and high PPV were obtained (sensitivity = 0.60, PPV = 0.80) while training on I2 and testing on I1, which indicate that a lot of false negatives are present, and high sensitivity and low PPV were obtained (sensitivity = 0.89, PPV = 0.47) while training on I1 and testing on I2, which indicates a lot of false positives are present. For training and testing on the same institution, both sensitivity and PPV were greater than or equal to 0.65 (sensitivity = 0.79, PPV = 0.69 for I1) and (sensitivity = 0.84, PPV = 0.72 for I2). Similar to training/testing on the same institution, high sensitivity and PPV were obtained while training on both institutions (sensitivity = 0.65, PPV = 0.81 for I1) and (sensitivity = 0.80, PPV = 0.75 for I2). Full results are using sensitivity and PPV are reported in Tables III and IV, respectively.

The evaluation of segmentation performance for the model trained on randomly selected 24 cases from Henry Ford Hospital (I2) and tested on both remaining 22 cases from same institution and unseen 22 cases from MD Anderson Cancer Center (I1), is similar to our earlier findings. Dice coefficients of 0.76 for patients tested in I2 and 0.68 for patients tested in I1 were obtained. Full results for this experiment are reported in Table V.

4. DISCUSSION

In this study, we investigated the effect of using multi-institutional data for the GBM segmentation by CNNs. We have experimentally demonstrated that segmentation accuracy highly depends on the training set. Specifically, we found that:

-Training on data from a different institution may dramatically deteriorate performance of the model
-Given data from the same institution where the model will be applied, amending it with data from a different institution can result in an improvement in performance of the model for some class labels for one institution. However, this conclusion does not hold for segmentation of the entire tumor.

While determining the specific reasons for the decreased performance when trained with data from a different institution is beyond the scope of this study, we hypothesize that one of the reasons is likely to be scanner parameters and acquisition parameters as we found that those differed significantly between the two institutions. Specifically, in MD Anderson Cancer Center the magnetic field strength for most of the cases is 1.5 Tesla and slice thickness is 5 mm. On the other hand, for Henry Ford Hospital the magnetic field strength for most of the cases is 3 Tesla and the slice thickness varies from one case to another between 3 and 5 mm.

We applied minimal preprocessing to the available sequences. The only preprocessing step that we applied was the registration among the sequences. Thus, the trained models are reflective of the different characteristics related to each institution. Deep learning methodologies can learn the inherent variations present in the raw data[26] and find an effective representation of the underlying structure of the data. However, our analyses show that CNNs are strongly affected by variations present in the data source. It is possible that additional preprocessing techniques that intend to unify the appearance and characteristics of the data could reduce the issues of interinstitutional training and testing in CNN. However, further investigations are required to draw any conclusion about this. This study aims to determine the extent of the effect of interinstitutional training rather than evaluate the effectiveness of different preprocessing methods to address this issue.

The results of our study in terms of the overall segmentation performance fit within a range of performance reported by other investigators. While recent results report somewhat better segmentation performance (Dice coefficient above 0.8), this is expected due to a low number of training cases and no preprocessing applied in our study. Dice coefficients above 0.8 for segmentation of the entire tumor (Class 1 in our case) were reported in several studies (comparisons of several of these techniques are presented in Ref. [6,7]), however, preprocessing steps of normalization and interpolation were carried out in those. Despite these factors, we obtained slightly better results than those in a recent study with minimal preprocessing,[27] where Dice coefficients of 0.72 for whole tumor and 0.57 for the active tumor were reported for training with 32 patients in BRATS dataset. When considering other alternative approaches for segmentation such as random forest-based segmentation[18] proposed by our group, we achieve very similar performance of average Dice coefficient 0.75 was obtained for Class 1, followed by 0.63 for Class 2 while training and testing within the same institution.

TABLE II. Average Hausdorff Distance (AHD) for different classes for different types of train-test combinations using 10-fold cross-validation.

	Class 1		Class 2		Class 3		Class 4		Class 5	
Train	Test on I1	Test on I2	Test on I1	Test on I2	Test on I1	Test on I2	Test on I1	Test on I2	Test on I1	Test on I2
Same institution	2.58	1.24	4.82	3.81	5.24	2.81	3.69	2.69	3.86	2.07
Different institutions	4.98	1.36	7.96	3.90	8.02	5.73	5.90	2.98	6.42	2.22
Both institutions	4.13	1.46	6.93	4.36	6.86	6.40	5.05	2.77	5.55	2.46

TABLE III. Sensitivity for different classes for different types of train-test combinations using 10-fold cross-validation.

	Class 1		Class 2		Class 3		Class 4		Class 5	
Train	Test on I1	Test on I2	Test on I1	Test on I2	Test on I1	Test on I2	Test on I1	Test on I2	Test on I1	Test on I2
Same institution	0.79 ± 0.17	0.84 ± 0.10	0.24 ± 0.11	0.20 ± 0.10	0.12 ± 0.10	0.09 ± 0.07	0.15 ± 0.07	0.11 ± 0.07	0.28 ± 0.16	0.44 ± 0.16
Different institutions	0.61 ± 0.21	0.89 ± 0.12	0.18 ± 0.10	0.22 ± 0.13	0.09 ± 0.07	0.10 ± 0.10	0.13 ± 0.07	0.09 ± 0.08	0.21 ± 0.13	0.49 ± 0.20
Both institutions	0.65 ± 0.22	0.80 ± 0.13	0.20 ± 0.10	0.19 ± 0.10	0.10 ± 0.07	0.08 ± 0.07	0.14 ± 0.08	0.11 ± 0.06	0.23 ± 0.14	0.42 ± 0.16

TABLE IV. Positive Predictive Value (PPV) for different classes for different types of train-test combinations using 10-fold cross-validation.

	Class 1		Class 2		Class 3		Class 4		Class 5	
Train	Test on I1	Test on I2	Test on I1	Test on I2	Test on I1	Test on I2	Test on I1	Test on I2	Test on I1	Test on I2
Same institution	0.69 ± 0.21	0.72 ± 0.16	0.79 ± 0.22	0.84 ± 0.17	0.85 ± 0.23	0.91 ± 0.20	0.55 ± 0.23	0.53 ± 0.20	0.67 ± 0.21	0.71 ± 0.14
Different institutions	0.81 ± 0.14	0.48 ± 0.21	0.87 ± 0.13	0.55 ± 0.23	0.95 ± 0.07	0.60 ± 0.29	0.70 ± 0.17	0.29 ± 0.24	0.80 ± 0.13	0.47 ± 0.19
Both institutions	0.81 ± 0.13	0.75 ± 0.17	0.89 ± 0.12	0.86 ± 0.18	0.94 ± 0.15	0.91 ± 0.20	0.71 ± 0.16	0.58 ± .0.21	0.80 ± 0.14	0.74 ± 0.15

1156

TABLE V. Average Dice coefficients for different classes for the 22 unseen cases from MD Anderson Cancer Center (I1) and Henry Ford Hospital (I2) using the model trained on randomly selected 24 cases from Henry Ford Hospital (I2).

	Class 1	Class 2	Class 3	Class 4	Class 5
I1	0.68 ± 0.20	0.30 ± 0.15	0.17 ± 0.12	0.20 ± 0.11	0.32 ± 0.18
I2	0.76 ± 0.12	0.35 ± 0.11	0.17 ± 0.15	0.18 ± 0.10	0.46 ± 0.155

The results for Class 3, 4, and 5 are much worse than compared to those of Classes 1 and 2. The median proportions of all the experiments conducted of classes (0 and 2–5), were as almost 80% to 20%. The primary reason is a higher difficulty[3] of segmenting individual components of the tumor as well as a significantly lower number of training patches available for the small components. For individual classes, the proportions of patches were less than 7% and we had less number of patches for individual classes to train with. Our performance to segment necrosis (Class 3) is close to the segmentation performance reported in Ref. [3] by a modified three-dimensional CNN segmentation technique.

Our study has some limitations in terms of the number of cases per institution and the number of institutions that we used. We have used only 22 cases for each institution because that was the maximum number of cases available from MD Anderson Cancer Center with the three sequences required for training and based on this we randomly selected 22 cases from Henry Ford Hospital. We could not include data of any other institution because all of them have a number of cases less than 22. Please note, however, that (a) this offered a sufficient variability among cases to obtain a reasonable segmentation performance and (b) we were still able to use 3–6 million training examples (patches), bringing this number close to a typical number of data used for this kind of training. Moreover, our study represents a scenario where the number of cases for learning is limited, proportion of positive examples (tumor in our case) is significantly lower than that of negative examples, and the data sources vary in terms of protocols. This is quite typical in medical image segmentation problems. Evaluation of the performance of CNN in this typical context is important and we found that performance of CNN is susceptible to the variations in the source of the data.

It is to be noted that all our tumor annotations were validated/corrected by a neuroradiology fellow after the initial annotation by a researcher. Thus, the ground truth is generated jointly by two readers. Therefore, the CNN models trained on this ground truth were biased with the segmentation techniques of two readers. However, this does not affect our conclusion that the training with same institutions improves the segmentation performance. However, we cannot account for interobserver variability in the work due to the absence of multiple ground truths generated by several readers. For generalizability of our results in the presence of interobserver variability, this study needs to be done with significantly more number of readers, and this can be thought of as an area for future research.

In summary, we established that the choice of the training set in a multi-institutional setting has a profound impact on the resulting performance of the model and significant care should be taken to account for this issue. Determination of the precise reasons behind this effect is another area of future research. Future work could also focus on developing methods that can alleviate this effect. Specifically, various methods have been proposed to bring images from different scanners into the same framework including the nonuniformity correction algorithm N4ITK.[28] Repeated studies such as the one presented here could be conducted after application of each of the algorithms to establish their effect on final segmentation performance by CNNs. Even though our study was based on GBM patients, multi-institutional dataset is used in other diseases as well. Hence, our observations may hold significance for different diseases. The research area of fully automated radiomics/radiogenomics may also benefit from our observations.

5. CONCLUSION

We found that the institutional bias affects the performance of the CNNs used to segment GBM from brain tumor patients. Therefore, the choice of training dataset has a strong effect on the segmentation results in a multi-institutional setting.

CONFLICTS OF INTEREST

The authors declare that they do not have any conflict of interest for this study.

[a]Author to whom corrrespondence should be addressed. Electronic mail: maciej.mazurowski@duke.edu; Telephone: +1 (919) 684 1466; Fax: +1 (919) 684 1491.

REFERENCES

1. Gordillo N, Montseny E, Sobrevilla P. State of the art survey on MRI brain tumor segmentation. *Magn Reson Imaging*. 2013;31:1426–1438.
2. Bauer S, Wiest R, Nolte L-P, Reyes M. A survey of MRI-based medical image analysis for brain tumor studies. *Phys Med Biol*. 2013;58:R97.
3. Kamnitsas K, Ferrante E, Parisot S, et al. "DeepMedic for Brain Tumor Segmentation." pp. 138-149; 2017.
4. Zhao L, Jia K. "Deep feature learning with discrimination mechanism for brain tumor segmentation and diagnosis." pp. 306-309; 2016.
5. Dvorak P, Menze B. "Structured prediction with convolutional neural networks for multimodal brain tumor segmentation." Proceeding of the Multimodal Brain Tumor Image Segmentation Challenge, pp. 13-24; 2015.
6. Havaei M, Davy A, Warde-Farley D, et al. Brain tumor segmentation with deep neural networks. *Med Image Anal*. 2017;35:18–31.

7. Akkus Z, Galimzianova A, Hoogi A, Rubin DL, Erickson BJ. Deep learning for brain MRI segmentation: state of the art and future directions. *J Digit Imaging*. 2017;30:449–459.
8. Pereira S, Pinto A, Alves V, Silva CA. Brain tumor segmentation using convolutional neural networks in MRI images. *IEEE Trans Med Imaging*. 2016;35:1240–1251.
9. Havaei M, Davy A, Warde-Farley D, et al. Brain tumor segmentation with deep neural networks. *Med Image Anal*. 2016;35:18–31.
10. Urban G, Bendszus M, Hamprecht F, Kleesiek J. Multi-modal brain tumor segmentation using deep convolutional neural networks. *MICCAI BraTS (Brain Tumor Segmentation) Challenge*. Proceedings, winning contribution, pp. 31-35; 2014.
11. Zikic D, Ioannou Y, Brown M, Criminisi A. "Segmentation of brain tumor tissues with convolutional neural networks," Proceedings MICCAI-BRATS, pp. 36-39; 2014.
12. Kamnitsas K, Ledig C, Newcombe VFJ, et al. Efficient multi-scale 3D CNN with fully connected CRF for accurate brain lesion segmentation. *Med Image Anal*. 2017;36:61–78.
13. Yi D, Zhou M, Chen Z, Gevaert O. "3-D Convolutional Neural Networks for Glioblastoma Segmentation," arXiv preprint arXiv:1611.04534; 2016.
14. Pereira S, Pinto A, Alves V, Silva CA. "Deep convolutional neural networks for the segmentation of gliomas in multi-sequence MRI." pp. 131-143.
15. Menze BH, Jakab A, Bauer S, et al. The multimodal brain tumor image segmentation benchmark (BRATS). *IEEE Trans Med Imaging*. 2015;34:1993–2024.
16. Czarnek NM, Clark KL, Peters KB, Mazurowski MA. Algorithmic three-dimensional analysis of tumor shape in MRI improves prognosis of survival in glioblastoma: a multi-institutional study. *J Neurooncol*. 2017;132:55–62.
17. Mazurowski MA, Czarnek NM, Collins LM, Peters KB, Clark K. "Predicting outcomes in glioblastoma patients using computerized analysis of tumor shape: preliminary data." pp. 97852T-97852T-6.
18. Zhang J, Barboriak DP, Hobbs H, Mazurowski MA. A fully automatic extraction of magnetic resonance image features in glioblastoma patients. *Med Phys*. 2014;41:042301.
19. Lambin P, Rios-Velazquez E, Leijenaar R, et al. Radiomics: extracting more information from medical images using advanced feature analysis. *Eur J Cancer*. 2010;48:441–446.
20. Mazurowski MA. Radiogenomics: what it is and why it is important. *J Am Coll Radiol*. 2015;12:862–866.
21. Gevaert O, Xu J, Hoang CD, et al. Non–small cell lung cancer: identifying prognostic imaging biomarkers by leveraging public gene expression microarray data—methods and preliminary results. *Radiology*. 2012;264:387–396.
22. Greve DN, Fischl B. Accurate and robust brain image alignment using boundary-based registration. *NeuroImage*. 2009;48:63–72.
23. Nair V, Hinton GE. "Rectified linear units improve restricted boltzmann machines." pp. 807-814; 2010.
24. Dice LR. Measures of the amount of ecologic association between species. *Ecology*. 1945;26:297–302.
25. Dubuisson MP, Jain AK. "A modified Hausdorff distance for object matching." pp. 566-568.
26. LeCun Y, Bengio Y, Hinton G. Deep learning. *Nature*. 2015;521:436–444.
27. Pandian B, Boyle J, Orringer DA. "Multimodal tumor segmentation with 3D convolutional neural networks". Proceedings of MICCAI-BRATS Workshop, pp. 49-52; 2016.
28. Tustison NJ, Avants BB, Cook PA, et al. N4ITK: improved N3 bias correction. *IEEE Trans Med Imaging*. 2010;29:1310–1320.