Glioma Segmentation and a Simple Accurate Model for Overall Survival Prediction

Evan Gates[1], J. Gregory Pauloski[1], Dawid Schellingerhout[2], and David Fuentes[1(✉)]

[1] Department of Imaging Physics, University of Texas MD Anderson Cancer Center, Houston, TX, USA
{EGates1,DTfuentes}@mdanderson.org
[2] Department of Cancer Systems Imaging and Diagnostic Radiology, University of Texas MD Anderson Cancer Center, Houston, TX, USA

Abstract. Brain tumor segmentation is a challenging task necessary for quantitative tumor analysis and diagnosis. We apply a multi-scale convolutional neural network based on the DeepMedic to segment glioma subvolumes provided in the 2018 MICCAI Brain Tumor Segmentation Challenge. We go on to extract intensity and shape features from the images and cross-validate machine learning models to predict overall survival. Using only the mean FLAIR intensity, nonenhancing tumor volume, and patient age we are able to predict patient overall survival with reasonable accuracy.

Keywords: Glioblastoma · Segmentation · Neural network · Quantitative imaging

1 Introduction

Gliomas are highly malignant primary brain tumors that carry a dismal median overall survival of 15 months for high grade tumors [1]. One characteristic that contributes to this poor survival is the substantial heterogeneity. Spatial heterogeneity within a tumor implicitly increases the chances that a therapy resistant tumor subpopulation exists and thus frequently indicates poor clinical prognosis [2]. Successful and automated detection of distinct subvolumes (enhancing, nonenhancing, and necrotic regions, etc.) is a key step in quantitative analysis towards patient risk stratification and computer aided diagnosis. In recent years, convolutional neural networks (CNNs) are the undisputed champions of biomedical segmentation tasks [3]. Quantitative measurements of these subvolumes are likely to provide insight into patient's prognosis.

In this work we use a multi-scale convolutional neural network to segment glioma sub-volumes in multi-contrast MRI images. We go on to extract shape and intensity features from the sub-volumes to predict patient overall survival.

A. Crimi et al. (Eds.): BrainLes 2018, LNCS 11384, pp. 476–484, 2019.
https://doi.org/10.1007/978-3-030-11726-9_42

Results on the 2018 MICCAI Brain Tumor Segmentation (BraTS) Challenge [4–7] are provided. Final challenge rankings relative to other contest entries are available online [8].

2 Segmentation

2.1 Network Structure

Data Preprocessing. The BraTS 2018 Training set contains 285 multi-contrast MRI (T1, T1ce, T2, FLAIR) scans of high and low-grade gliomas. 75 of the 285 patients are labeled low-grade (LGG) and the remaining are high-grade (HGG). The imaging data is brain extracted, registered, and resampled to 1 mm isotropic voxel size. Each subject has a ground truth segmentation with four labels, non-tumor (label 0), necrotic and nonenhancing tumor core (label 1), peritumoral edema (label 2), and Gadolinium-enhancing tumor (label 4). The BraTS 2018 Validation set contains a mix of 66 HGG and LGG patients equivalently pre-processed and does not have ground truth segmentations.

All MRI scans were normalized by subtracting the mean intensity and dividing by the standard deviation. A binary brain mask for each patient was also created using the T1 scan, and this mask is used by the CNN to focus sampling on only the brain. The same preproccesing steps were also applied to the validation data set before segmenting.

Convolutional Neural Network. We used a 3-dimensional CNN built using the DeepMedic architecture created by Kamnitsas et al. [9]. DeepMedic has consistently produced high performing image segmentations in previous BraTS challenges. Sampling was used to produce image sub-volumes of size 37^3, and an equal number of sub-volumes centered on the foreground and background was taken to reduce class imbalance. Our CNN implementation contains three pathways consisting of eleven convolutional layers each. The pathways include one normal resolution and two downsampled where one was downsampled by a factor of 3^3 and the other by 5^3. The first seven layers use a 3^3 kernel with 30 to 50 features per layer. After the first seven layers, the downsampled pathways are upsampled to match the normal resolution pathway and all three pathways are concatenated. The concatenated features are then fed into two fully connected layers with 250 features and a kernel size of 3^3 and 1^3 respectively. The final layer is a fully connected layer with kernel of size 1^3 and four features. Dropout rates of 50% were used on the final two layers to prevent overfitting. The four features in the last fully connected layer are the output of the network and represent binary masks for each of the four segmentation labels.

Initially the CNN was trained on 80% of the BraTS 2018 Training Data and the remaining 20% was reserved for model validation. We tuned the batch size, learning rate, and optimizer until we found a set of parameters that gave the most accurate validation results and efficient use of our hardware. The final network was trained for 50 epochs with a batch size of 10, and the RMSprop optimizer

was used with an initial learning rate of 0.001 and lowered throughout training. Before performing inference on the validation set we retrained the network with 95% of the BraTS 2018 Training Data and 5% reserved for validation. This training on a Nvidia Kepler Titan 6 GB took 96 h and, using the trained CNN, we performed a full inference on the BraTS 2018 Validation Data to produce binary mask for each of the four segmentation labels.

2.2 Training and Validation Set Results

The segmentation results from the full inference on the validation set were uploaded to the BraTS Challenge portal where the Dice score, sensitivity, specificity, and Hausdorff distance were calculated. Results for the 66 patients in the validation set are shown in Table 1. We also performed a full inference on the 5% of training set cases which were excluded from the model training process, so that we can compare the performance of the model against the ground truth segmentations. This comparison is shown in Table 2. From these samples, we can see that the CNN classifies the overall tumor well but has greater difficulty classifying regions with a dense mix of lower and higher grades.

Table 1. Mean values for metrics from segmentations on the training and validation data sets.

Data set	Label	Dice	Sensitivity	Specificity	Hausdorff
Training	Enhancing tumor	0.7332	0.84265	0.99781	6.20545
	Whole tumor	0.89633	0.88636	0.99531	5.14866
	Tumor core	0.75292	0.73297	0.99833	8.47618
Validation	Enhancing tumor	0.67831	0.72923	0.99611	14.52297
	Whole tumor	0.80558	0.81374	0.98703	14.415
	Tumor core	0.6852	0.68018	0.99619	20.01745

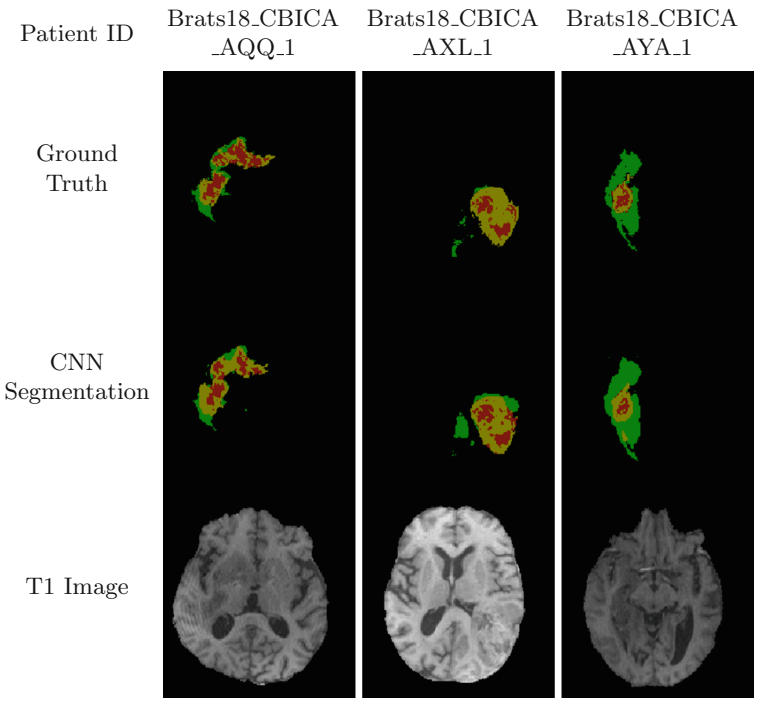
3 Survival Analysis

Of the 285 training cases, 163 cases had age (range 19–86 years) and overall survival data (range 5–1767 days) provided. We used this clinical data and features extracted from segmentations to predict the overall survival in days.

After determining the best model, we applied it to the 28 challenge validation set cases with patient age and gross total resection status.

The challenge assesses the predictions based on the accuracy: total number of cases correctly assigned a survival <10 months, between 10 and 15 months, and >15 months. The mean-square-error (MSE) is also used as a performance metric.

Table 2. Ground truth versus CNN segmentation samples from the validation data excluded from training. Red is nonenhancing tumor and necrosis (tumor core), green is edema, and yellow is enhancing tumor. The T1 weighted image for each patient is shown for reference.



3.1 Image Processing

The format of the provided imaging data is described in Sect. 2.1. For the survival task, we further pre-processed the data by normalizing based on reference tissue intensities. Creating a consistent intensity scale between patients allows images features to discriminate short and long survival patients. Note, this is different than the normalization used in the segmentation task where each image had mean zero and standard deviation one. To apply this normalization, we placed small regions of interest for each patient in the gray matter (GM) of the lentiform nucleus, the cerebrospinal fluid (CSF) of the ventricles, and the normal appearing white matter (WM). Using the mean intensity for a pair of reference tissues, each voxel in the image was linearly scaled to map the mean intensities to 0 and 1 respectively. For example, in the FLAIR image each voxel value x was transformed according to

$$y_{CSF/WM} = \frac{x - \overline{CSF}}{\overline{WM} - \overline{CSF}}$$

For the FLAIR image CSF and WM were chosen because they were the darkest and brightest reference tissues respectively. For sequences T1 and FLAIR we normalized using the CSF/WM pair, for T2 we used the WM/CSF pair, and for the T1 contrast enhanced image we used a CSF/GM pair. This procedure is similar to other methods presented in the literature [10]. Although we performed the normalization semi-automatically with manually placed ROIs, this procedure can be performed fully automated using brain tissue segmentation software applied to the non-tumor regions.

3.2 Image Features

To predict patient overall survival we calculated image features for each of the available image sequences and segmentation labels from Sect. 2. We also computed the union of the three regions (nonenhancing, enhancing, and edema) to generate a whole-tumor ROI for each patient. For each region (enhancing, nonenhancing, edema, and whole tumor) we computed the mean intensity of that region for each image. (T1, T1 contrast enhances, T2, FLAIR) as well as the volume using the *Pyradiomics* software package [11]. So, in total 20 features (16 means and 4 volumes) were used for predicting overall survival.

We experimented with features quantifying higher order histogram statistics (quantiles, skewness, etc) and complex shape descriptors (i.e. flatness). However, we found these features did not improve the performance of predictive modeling beyond using just mean values. Similarly, we quantified image texture using gray level co-occurrence matrices and gray level run length matrices, and nearest gray tone difference matrices [12] but again found that including these features did not substantially increase model performance. Since these higher-order features are less robust to variability in the underlying image data and segmentation, we chose to consider only mean intensity and volume features in our final analysis.

3.3 Survival Task

An overview of our model development approach to predict survival is shown in Fig. 1. A family of models was considered with distinct permutations of variable selection methods and machine learning prediction algorithm. The best model was used to make predictions on the provided validation data. Modeling for the survival task was implemented in R version 3.4.0.

We partitioned the training data into 80% training and 20% testing data with an approximately equal proportion of short, medium, and long survivors in each set. Using the 80% partition, we performed variable selection and trained several classes of predictive models including linear models, neural networks, and random forests using leave-one-out cross validation. We selected the model with the highest Pearson correlation (R^2) between predicted and observed overall survival within the cross validation and made predictions on the testing set to see how well the model generalized.

For feature selection we consecutively applied univariate, multivariate, and step-wise feature elimination. After each selection step the resulting variables

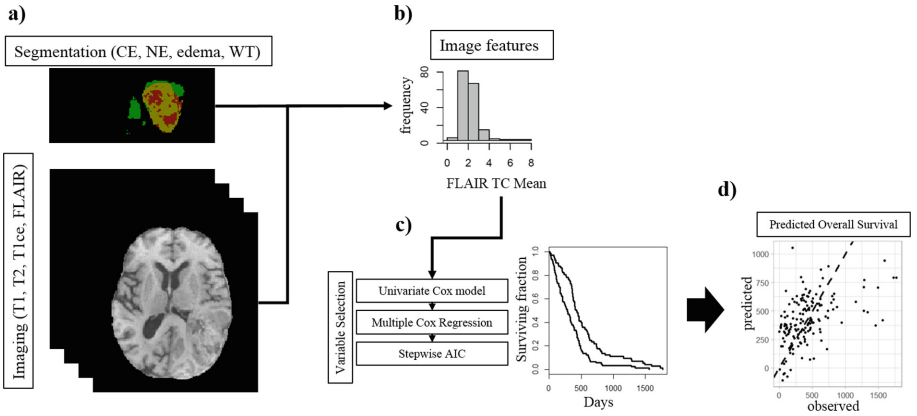


Fig. 1. Flowchart depicting the modeling process and variable selection methods for predicting overall survival. First the imaging data and computed segmentations **(a)** are used to extract mean intensity features and volumes. Panel **(b)** shows a histogram of the mean FLAIR intensity over the tumor core (TC) region. Several variable selection methods based on the cox model are used to generate input sets for predictive models **(c)**. Features were first tested for significant association with overall survival using univariate cox models and discarding non-significant features. The set of univariate significant features was further reduced by constructing a multivariate cox model and again eliminating redundant (non-significant features), followed by stepwise AIC. The remaining image features and the patient's age, an important clinical factor, were used to predict overall survival **(d)**. Before the variable selection 20% of the training data was held out for as an independent testing and the remaining 80% is used to select the best variable and model combination.

were stored as a possible set of inputs to predictive models. First, we used a Cox model to individually determine which image features were significantly associated with overall survival. Any feature with $p > 0.05$ for the Wald-test was discarded. Next, the remaining features were fed into a multivariate Cox model to reduce redundancy. Features with $p < 0.05$ in the multivariate Cox model were retained. Lastly, we further reduced the set of inputs using step-wise elimination based on Akaike Information Criteria (AIC) [13]. Starting with all variables, the stepwise AIC algorithm eliminates or replaces variables one at a time to maximize the AIC.

In addition, we applied the Boruta method [14] to select variables predictive of overall survival. The Boruta method is based on variable importance from the random forest algorithm, which has traditionally been a top performing machine leaning model.

Each variable subset was used to train several candidate models for predicting overall survival. We tested a linear model, random forest, and neural network and assessed the average cross-validation accuracy of each. After selecting the best model and variable combination, we trained a final model on all the training data, made predictions on the challenge-provided validation set, and compared

the performance to the leave-one-out cross validation. In particular, we checked for substantially decreased performance on the test data that would indicate model over-fitting.

3.4 Results

Landmark normalization was successfully applied to all patients. One case had poor fluid suppression on the FLAIR image and could not be effectively normalized. This case was excluded from the training.

Among the mean intensities for each image over each region, the region volumes, and the patient age, we found five features were significantly associated with overall survival in the cox model. They are: mean FLAIR intensity over the nonenhancing and necrotic region, mean T1ce intensity over the whole tumor, the volumes of the nonenhancing and enhancing regions, and age. With these variables input into a multivariate Cox model only age, the FLAIR nonenhancing mean, and nonenhancing volume were independently significant. Applying stepwise AIC did not change the variable selections any further.

Among the candidate models we tested (random forest, neural network, linear model) the linear model performed best with $R^2 = 0.134$ and mean-square-error 114994 using the three inputs selected by the multivariate cox model. With the same model parameters fit to all 162 evaluable challenge cases the model to predict overall survival is given by.

$$\text{Survival} = 926.8 - 10.5 \cdot \text{Age} + 91.6 \cdot \text{FTCM} - 55.1 \cdot \text{TCV}$$

where Age is the patient's age in years, FTCM is the "FLAIR Tumor Core Mean" value on the landmark normalized scale, and TCV is the "Tumor Core Volume" in units of $\text{mm}^3/10000$ consisting of nonenhancing and necrotic areas. This volume scaling makes the range of values comparable to the other features. Surprisingly, this simple linear model performed substantially better on the testing data and on the challenge validation dataset. This strongly suggests the model is not over fitting the data. The metrics are provided in Table 3.

Table 3. Performance metrics for our linear model on the training data: (80% of 163 provided cases), testing data (20% of 163 provided cases), and validation data (26 cases without known survival). The Pearson R^2 for the validation data is not provided.

	R^2, predicted vs observed	Accuracy	MSE
Training data	0.134	44.5%	114994
Testing data	0.399	38.2%	55193
Validation data	-	53.6%	87998

4 Discussion

Brain tumor segmentation and prediction of overall survival are both challenging tasks. Despite good results, our segmentation model did not perform as well as the implementation of DeepMedic by Kamnitsas et al. that won the BraTS 2017 challenge [15]. Their model achieved better segmentation results by averaging results across an ensemble of six different models. The single model we used is not as robust as their ensemble method but provides satisfactory results without the high computational cost.

In the task of predicting patients as short, medium, or long survivors we achieved a validation accuracy of 54% with a MSE of 87998. In the training data the most frequent class is short survivors at 65 of 163 (39.9%) which means our models are performing better than chance. The root mean square error for continuous prediction is on the order of 300 days, which is comparable to the range seen among all patients. Overall survival is impacted by several factors, including age, treatment, and performance status (not provided) and the accuracy and MSE reflects the complexity of this task even when some variables are controlled for.

We were able to produce good results using two highly primitive image measurements (mean intensity and volume) and a linear regression model. Although vast numbers of higher-order texture features and nonlinear models are commonly employed to mine imaging data, we found they were not useful in predicting overall survival for this task. We suspect this is because these features are more sensitive to tumor segmentation (and segmentation error) as well as other variations in image quality and processing. Since predicting overall survival is already a highly uncertain task, it is easy for models to over-fit the higher order features. In other words, the simple and robust features more easy to generalize.

Our best performing model only included intensity information from one of the four magnetic resonance sequences available (FLAIR) and only one of the four segmentation labels used to extract features (enhancing tumor, tumor core consisting of nonenhancing tumor and necrosis, edema, and whole tumor). This may have happened for a few reasons: While the available image types (T1, T2, etc) contain different kinds of information about the tumors, there was a lot of variability between images of the same type from different patients. This intra-sequence variability reduces the impedes the models ability to predict overall survival based on the complementary nature of the different image contrasts.

5 Conclusion

We found we could segment glioma tumors with high accuracy using a multi-scale convolutional neural net. Using these segmentations and simple image features we were able to predict overall survival with reasonable accuracy.

References

1. Stupp, R., et al.: The European organisation for research and treatment of cancer brain tumor and radiotherapy groups, and "the national cancer institute of canada clinical trials group". radiotherapy plus concomitant and adjuvant temozolomide for glioblastoma. N. Engl. J. Med. **352**(10), 987–996 (2005)
2. Shipitsin, M., et al.: Molecular definition of breast tumor heterogeneity. Cancer cell **11**(3), 259–273 (2007)
3. Ronneberger, O., Fischer, P., Brox, T.: U-Net: convolutional networks for biomedical image segmentation. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) MICCAI 2015. LNCS, vol. 9351, pp. 234–241. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-24574-4_28
4. Menze, B.H., et al.: The multimodal brain tumor image segmentation benchmark (BRATS). IEEE Trans. Med. Imaging **34**(10), 1993–2024 (2015)
5. Bakas, S., et al.: Advancing the cancer genome atlas glioma MRI collections with expert segmentation labels and radiomic features. Sci. Data **4**, 170117 (2017)
6. Bakas, S., et al.: Segmentation labels and radiomic features for the pre-operative scans of the TCGA-GBM collection, The Cancer Imaging Archive (2017). https://doi.org/10.7937/K9/TCIA.2017.KLXWJJ1Q
7. Bakas, S., et al.: Segmentation labels and radiomic features for the pre-operative scans of the TCGA-LGG collection, The Cancer Imaging Archive (2017). https://doi.org/10.7937/K9/TCIA.2017.GJQ7R0EF
8. Bakas, S., Reyes, M., et al.: Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the BRATS challenge. arXiv preprint arXiv:1811.02629 (2018)
9. Kamnitsas, K., et al.: Efficient multi-scale 3D CNN with fully connected CRF for accurate brain lesion segmentation. Med. Image Anal. **36**, 61–78 (2017)
10. Leung, K.K., et al.: Alzheimer's disease neuroimaging initiative. robust atrophy rate measurement in alzheimer's disease using multi-site serial MRI: tissue-specific intensity normalization and parameter selection. NeuroImage **50**(2), 516–523 (2010)
11. van Griethuysen, J.J.M., et al.: Computational radiomics system to decode the radiographic phenotype. Cancer Res. **77**(21), e104 LP–e107 (2017)
12. Haralick, R., Shanmugan, K., Dinstein, I.: Textural features for image classification (1973)
13. Akaike, H.: A new look at the statistical model identification. IEEE Trans. Autom. Control **19**(6), 716–723 (1974)
14. Kursa, M.B., Rudnicki, W.R., et al.: Feature selection with the Boruta package. J. Stat. Softw. **36**(11), 1–13 (2010)
15. Kamnitsas, K., et al.: Ensembles of multiple models and architectures for robust brain tumour segmentation. CoRR, abs/1711.01468 (2017)