Towards Uncertainty-Assisted Brain Tumor Segmentation and Survival Prediction

Alain Jungo[1](✉), Richard McKinley[2], Raphael Meier[1], Urspeter Knecht[2], Luis Vera[3], Julián Pérez-Beteta[3], David Molina-García[3], Víctor M. Pérez-García[3], Roland Wiest[2], and Mauricio Reyes[1]

[1] Institute for Surgical Technology and Biomechanics,
University of Bern, Bern, Switzerland
alain.jungo@istb.unibe.ch

[2] Support Center for Advanced Neuroimaging,
Institute for Diagnostic and Interventional Neuroradiology,
University Hospital Inselspital and University of Bern, Bern, Switzerland

[3] Mathematical Oncology Laboratory, Universidad de Castilla-La Mancha,
Ciudad Real, Spain

Abstract. Uncertainty measures of medical image analysis technologies, such as deep learning, are expected to facilitate their clinical acceptance and synergies with human expertise. Therefore, we propose a full-resolution residual convolutional neural network (FRRN) for brain tumor segmentation and examine the principle of Monte Carlo (MC) Dropout for uncertainty quantification by focusing on the Dropout position and rate. We further feed the resulting brain tumor segmentation into a survival prediction model, which is built on age and a subset of 26 image-derived geometrical features such as volume, volume ratios, surface, surface irregularity and statistics of the enhancing tumor rim width. The results show comparable segmentation performance between MC Dropout models and a standard weight scaling Dropout model. A qualitative evaluation further suggests that informative uncertainty can be obtained by applying MC Dropout after each convolution layer. For survival prediction, results suggest only using few features besides age. In the BraTS17 challenge, our method achieved the 2nd place in the survival task and completed the segmentation task in the 3rd best-performing cluster of statistically different approaches.

Keywords: Deep learning · Brain tumor segmentation
Uncertainty estimation · Survival prediction

1 Introduction

Over the past years, large improvements could be observed in brain tumor segmentation. This is partly due to the adoption of the fast-evolving deep learning approaches from the field of computer vision. An even more important reason for the recent advances is the availability of public datasets and online

A. Crimi et al. (Eds.): BrainLes 2017, LNCS 10670, pp. 474–485, 2018.
https://doi.org/10.1007/978-3-319-75238-9_40

benchmarks [15]. This progress has later guided research to focus on optimizing model architectures for achieving high segmentation performance. However, as the robustness of these systems still requires expert monitoring of results, clinical applications such as radiological and high-throughput data analysis would benefit greatly from additional uncertainty information along with a good segmentation performance. Information on the segmentation uncertainty can first leverage the trust of users on such automated segmentation systems, but it could be also used to e.g. guide an operator in making manual corrections to the automatic segmentation results. In this work, we thus focus on the largely unexplored aspect of quantifying model uncertainty in the context of brain tumor segmentation. Existing work of uncertainty in brain tumor segmentation includes a perturbation-based approach for conditional random fields [1,14] and a level set-based method defined via a Gaussian Process [13]. The limitations of these techniques are their lack of transferability to neural networks and their restriction to quantify uncertainty of a specific model only.

The segmentation of brain tumor compartments can be part of a radiomic process, where features are extracted from the segmented tumor and used in subsequent data mining [9]. A particularly important radiomic application is the survival prediction [20], where imaging features are used within a radiomics workflow to predict patient survival. Typically, features are handcrafted and include regional gray-level features (e.g., first- and second-order statistics) and morphological features (e.g., surface and volume [18]). In addition to imaging features, clinical features such as age or extent of resection may also be considered [5]. Geometrical features such as tumor surface irregularity or enhancing tumor heterogeneity have been reported as predictive biomarkers for patient survival [6,12,16]. However, these studies rely on manual or semi-automatic delineation of the tumor compartments.

The aim of this work is twofold. First, to explore uncertainty estimation in deep learning-based methods for brain tumor segmentation, and second, to predict survival from age and image-derived geometrical features of a segmented tumor shape. Therefore, as a baseline for the segmentation task, we adopt the a full-resolution residual network (FRRN) [17] architecture. Then, we incorporate the idea of Monte Carlo (MC) Dropout [8] to obtain model uncertainty. In an experiment, we examine the impact of different MC Dropout position strategies and compare the performance to the standard weight scaling Dropout [19]. For the survival prediction task, we use the resulting fully-automated segmentations to determine geometrical features and to build a predictive model thereof.

2 Methods

In this section, we present details of the approach subdivided for segmentation and survival prediction.

2.1 Segmentation

First, we present the employed convolutional neural network architecture. Second, we focus on the method used to quantify segmentation uncertainty.

FRRN Architecture. The adopted full-resolution residual network (FRRN) [17] is based on two streams; a residual stream and a pooling stream. The first one is responsible for maintaining a residual path between the network input and output. This has been shown to improve the gradient flow and thus training [10]. Moreover, the residual stream allows the network to carry information at full image resolution required for precise segmentation of the image details [17]. The second stream reduces the resolution by pooling operations before returning to original resolution by upsampling. Due to the reduced resolution, the filters on the pooling stream can capture contextual information. An important aspect of the architecture are the connections between pooling and residual streams. This enables the network to simultaneously combine both global and local image information [17].

We propose a full-resolution residual network architecture with four max-pooling/upsampling steps. The input consists of slices of all four image sequences (T1-weighted, T1-weighted post-contrast, T2-weighted, Fluid attenuation inversion recovery (FLAIR)), which we define as $I = [I_{T1}, I_{T1c}, I_{T2}, I_{FLAIR}] \in \mathbb{R}^{m \times n \times 4}$ where m and n are the in-plane resolution. The output is defined to be a segmentation mask for the corresponding input slice I, which contains the labels of the tumor compartments (edema, enhancing tumor and necrosis together with non-enhancing tumor) and the background, i.e. $\mathcal{C} = \{0, 1, 2, 4\}$. For each input I, the network determines the posterior probability distribution $p(Y \mid I)$ where $Y \in \mathcal{C}^{m \times n}$. Although the prediction is performed slice-wise, the formulation can be extended to subject volumes with $\mathbf{I} \in \mathbb{R}^{l \times m \times n \times 4}$ and $\mathbf{Y} \in \mathcal{C}^{l \times m \times n}$ (l being the slices).

The architecture of the network is depicted in Fig. 1. Figure 2 shows a detailed view of the residual units (RU) and full-resolution residual units (FRRU). Due to the anisotropy in the image resolution of the original data, we consider the slices from all three planes (axial (a), coronal (c) and sagittal (s)) by rotating the input volumes $\mathbf{I}$ to $\mathbf{I}_a$, $\mathbf{I}_c$ and $\mathbf{I}_s$ during training and testing. This results in three predictions $p(\mathbf{Y}_a \mid \mathbf{I}_a)$, $p(\mathbf{Y}_c \mid \mathbf{I}_c)$, $p(\mathbf{Y}_s \mid \mathbf{I}_s)$ per subject. In order to combine them, the three outputs are averaged to $p(\mathbf{Y} \mid \mathbf{I}) = 1/3 \sum_{j \in \{a,c,s\}} p'(\mathbf{Y}_j \mid \mathbf{I}_j)$ where p' denotes the posterior probabilities in the space of $\mathbf{I}$, before determining the maximizing class $\hat{y} = \arg\max_{c \in \mathcal{C}} p(y = c \mid I)$ for each voxel. Together with the volume-wise intensity normalization ($\mu = 0$, $\sigma = 1$), this combined prediction on all image planes had the largest impact on our validation set performance.

Uncertainty Estimation. As presented by Gal and Ghahramani [8], Dropout regularization can be interpreted as an approximation for Bayesian inference over the weights of the network. A fully Bayesian network requires applying Dropout after each convolution layer. Kendall et al. [11] showed that applying Dropout at

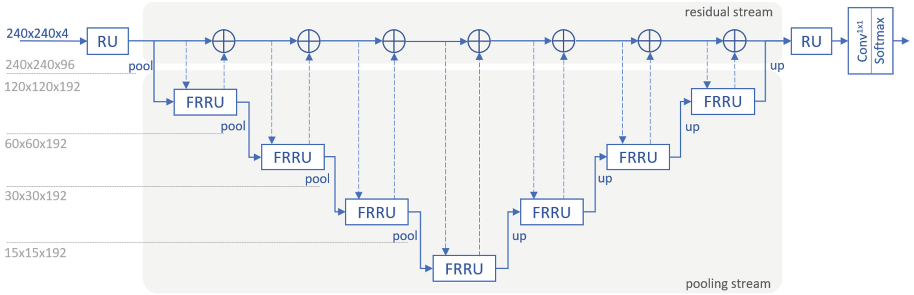
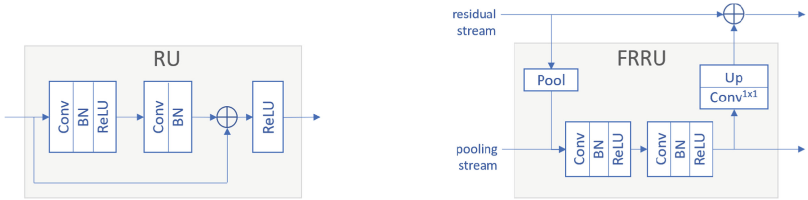


Fig. 1. Full-resolution residual network with four pooling steps. Dashed lines represent the exchange connections between the residual and pooling streams.



(a) (b)

Fig. 2. Detailed view of the units of the architecture in Fig. 1. (a) The residual unit (RU) including its residual connection (BN: Batch normalization, ReLU: Rectified Linear Unit). (b) The full-resolution residual unit (FRRU) where *pool* and *up* adapt to the pooling and residual stream, respectively. The 1×1 convolution aligns the number of feature channels among the streams. Unless specified differently, the convolution kernels are of size 3×3.

key positions of the network can be sufficient for semantic segmentation, and that it additionally favors training convergence. Following this observation, we select the positions after each pooling layer and before each upsampling operations as well as the position before and after the latter residual unit as key Dropout positions (Fig. 3). Hereinafter, these positions will be referred to as *core* and *end* Dropout positions.

The Dropouts are applied during training and test time. At test time, the Dropouts produce randomly sampled networks, which can be viewed as Monte Carlo samples over the posterior distribution $p(\mathbf{W} \mid \mathcal{I}, \mathcal{Y})$ of the model weights $\mathbf{W}$ (with subject dataset $\mathcal{I}$ and corresponding label set $\mathcal{Y}$). K network samples are used to produce one prediction with uncertainty estimation. The classification of one voxel is determined by the average of posterior probabilities $p(y \mid I) = \sum_{c \in \mathcal{C}} \left(\frac{1}{K} \sum_{k=1}^{K} p(y_k = c \mid I) \right)$ over K predictions. As described by Gal [7], the class uncertainty can be computed with the approximated predictive entropy $H \approx -\sum_{c \in \mathcal{C}} \left(\frac{1}{K} \sum_{k=1}^{K} p(y_k = c \mid I) \right) \log \left(\frac{1}{K} \sum_{k=1}^{K} p(y_k = c \mid I) \right)$.

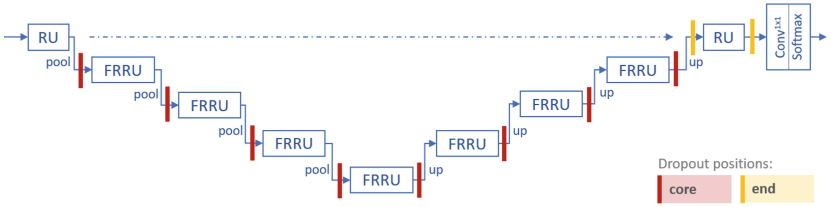


Fig. 3. Dropout positions *core* and *end* used for the different Dropout strategies.

2.2 Survival Prediction

Once the tumor compartments are segmented they serve as input for the overall survival prediction. Subsequent paragraphs elucidate the steps to build the survival prediction regression model.

Feature Extraction. As presented by Pérez-Beteta et al. [16], tumor geometry holds important information for survival prediction. Accordingly, we consider 26 geometrical features which include volumes (enhancing tumor (V_{ce}), necrosis (V_n), tumor core (V_c)), volume ratios (e.g., $\frac{V_c}{V_ce}$, $\frac{V_c}{V_n}$), surface, surface irregularity, maximal diameter as well as median, mean, quartiles and combinations of quartiles of the enhancing tumor rim width. In addition to the geometrical features, the subject's age is included.

Feature Selection. This step identifies the most important features before creating a prognostic model. We performed filtering with extensive cross-validation processes on the training set based on several information measurements (e.g. Gini impurity, variance reduction with respect to target attribute). It revealed that four seems to be the optimal number of features for our set. The four selected features are (listed according to their importance):

1. Age
2. Tumor core (enhancing tumor and necrotic tissue) surface
3. Surface irregularity (surface compared to sphere with equal volume)
4. 1st quartile of contrast-enhancing rim width

Survival Model. With the four selected features, we train a fully connected neural network with one hidden layer and linear activation function. Other prediction models such as SVM with RBF kernels, sparse grid or combinations of them were investigated but resulted in inferior performance.

3 Experiments and Results

In this section, we first focus on the segmentation performance of several MC Dropout models compared to a traditional Dropout model before we perform a qualitative evaluation of the obtained uncertainties. In a second experiment, we are interested in the survival prediction performance.

3.1 Segmentation

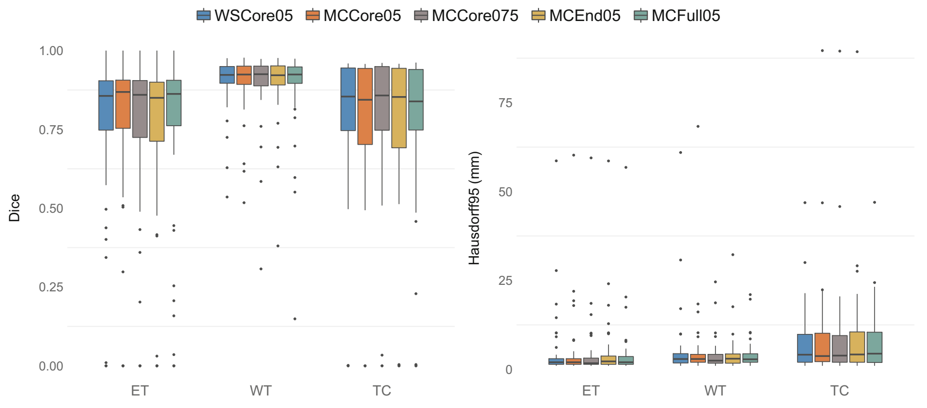
In order to examine MC Dropout, we compare four MC Dropout and one standard weight scaling (WS) Dropout [19] strategies. WS Dropout at the *core* positions is applied with a Dropout rate $p = 0.5$ (WSCore05). The four MC Dropout strategies are: Dropout at *core* positions with $p = 0.5$ (MCCore05) and with $p = 0.75$ (MCCore075), Dropout at *end* positions with $p = 0.5$ (MCEnd05) and Dropout after every convolution layer with $p = 0.5$ (MCFull05).

According to Kendall et al. [11] a minimum of approximately $K = 6$ Dropout Monte Carlo samples are required to improve segmentation performance (on the CamVid dataset) compared to an architecture where the Dropout weights are averaged during testing. For the MC Dropout models, we use a rather large $K = 20$. The reason is that compared to Kendall et al. [11], we are not only interested in an improved segmentation performance but also in exploiting the uncertainty comprised in the K predictions.

Table 1. Quantitative results of the comparison between the Dropout strategies WSCore05, MCCore05, MCCore075, MCEnd05, MCFull05 on the BraTS17 validation dataset (reported as mean ± standard deviation). Bold numbers highlight the best result for a given metric and tumor region (ET: enhancing tumor, WT: whole tumor, TC: tumor core).

	Model	ET	WT	TC
Dice	WSCore05	0.749 (±0.277)	**0.901** (±0.086)	**0.790** (±0.239)
	MCCore05	**0.756** (±0.275)	0.898 (±0.093)	0.775 (±0.245)
	MCCore075	0.730 (±0.290)	0.896 (±0.103)	0.776 (±0.243)
	MCEnd05	0.738 (±0.284)	0.894 (±0.114)	0.785 (±0.240)
	MCFull05	0.734 (±0.299)	0.884 (±0.141)	0.768 (±0.257)
Sensitivity	WSCore05	**0.800** (±0.273)	**0.900** (±0.130)	**0.760** (±0.263)
	MCCore05	0.775 (±0.273)	0.884 (±0.139)	0.721 (±0.270)
	MCCore075	0.783 (±0.263)	0.894 (±0.146)	0.740 (±0.267)
	MCEnd05	0.783 (±0.262)	0.882 (±0.154)	0.744 (±0.264)
	MCFull05	0.759 (±0.299)	0.857 (±0.177)	0.724 (±0.281)
Specificity	WSCore05	0.998 (±0.005)	0.995 (±0.004)	0.998 (±0.003)
	MCCore05	0.998 (±0.003)	0.996 (±0.005)	**0.999** (±0.003)
	MCCore075	0.998 (±0.005)	0.995 (±0.004)	0.998 (±0.003)
	MCEnd05	0.998 (±0.003)	0.996 (±0.004)	0.998 (±0.003)
	MCFull05	0.998 (±0.004)	**0.997** (±0.003)	0.998 (±0.004)
Hausdorff95(mm)	WSCore05	5.379 (±10.068)	5.409 (±9.710)	**7.487** (±8.935)
	MCCore05	5.025 (±10.098)	5.255 (±10.129)	8.842 (±15.023)
	MCCore075	5.425 (±9.812)	4.319 (±5.122)	8.909 (±14.292)
	MCEnd05	**4.671** (±9.600)	**4.059** (±4.349)	7.924 (±14.616)
	MCFull05	4.695 (±9.243)	4.216 (±4.166)	7.582 (±8.710)

The comparison of the approaches was performed on the 46 subjects of BraTS17 validation dataset [2–4,15]. All five models were trained on 265 randomly selected training subjects out of the 285 subjects available in the BraTS17 training dataset [2–4,15]. The remaining 20 training subjects were used for validation during training and model selection. Table 1 lists a summary of the achieved results for the five methods. Additionally, the distribution of the obtained Dice coefficients and Hausdorff (95th percentile) distances are presented in Fig. 4.



(a) (b)

Fig. 4. Boxplots for the Dice coefficient (a) and Hausdorff (95th percentile) distance (b) for the different Dropout strategies; weight scaling Dropout at *core* positions with Dropout rate $p = 0.5$ (WSCore05), Monte Carlo (MC) Dropout at *core* positions with $p = 0.5$ (MCCore05) and $p = 0.75$ (MCCore075), MC Dropout at *end* positions with $p = 0.5$ (MCEnd05) and MC Dropout after each convolution layer with $p = 0.5$ (MCFull05).

On the BraTS17 challenge dataset [2–4,15] with 146 subjects, the proposed method achieved the 7th rank in the segmentation task (results listed in Table 2). Furthermore, the method ranked third with regards to statistical differences among the approaches.

Table 2. BraTS17 challenge dataset results obtained by the WSCore05 model (reported as mean ± standard deviation, ET: enhancing tumor, WT: whole tumor, TC: tumor core).

	ET	WT	TC
Dice	0.670 (±0.312)	0.874 (±0.121)	0.736 (±0.304)
Hausdorff95(mm)	54.791 (±127.862)	8.825 (±15.550)	31.332 (±89.496)

As an additional output, the presented models produce uncertainty maps. A qualitative result is shown in Fig. 5 which depicts an exemplary case of an uncertainty map for each model along with the obtained segmentation. In contrast to the models with MC Dropout, the model applying weight scaling Dropout (WSCore05) does not perform Monte Carlo sampling. In this case, we determine the uncertainty through the entropy of the posterior probability distribution $H = -\sum_{c \in \mathcal{C}} p(y = c \mid I) \log p(y = c \mid I)$ for every voxel.

3.2 Survival Prediction

The evaluation of the survival prediction model was performed on the BraTS17 survival validation dataset [2–4, 15], which is a subset (33 out of 46 subjects) of the segmentation validation dataset. For this subset, the subject's age is provided with the data and is used as input for the prediction along with the computed segmentation results. Table 3 lists the results on the validation dataset (top row).

In the BraTS17 challenge, the presented model achieved the 2nd place in the survival task. The results achieved on the 95 subjects of the challenge dataset [2–4, 15] are presented in the bottom row of Table 3.

Table 3. Quantitative results of the survival prediction on the BraTS17 validation and challenge datasets.

Dataset	Accuracy	MSE	Median SE	Std SE
Validation	0.424	$245.7 \cdot 10^3$	$562.8 \cdot 10^3$	$540.7 \cdot 10^3$
Challenge	0.568	$213.0 \cdot 10^3$	$28.1 \cdot 10^3$	$662.6 \cdot 10^3$

4 Discussion

The evaluation of the validation dataset results in Table 1 reveals that the segmentation performance of all five models (WSCore05, MCCore05, MCCore075, MCEnd05, MCFull05) is comparable. Nevertheless, the best overall performance is achieved by the model with weight scaling Dropout (WSCore05). This is rather surprising since Kendall et al. [11] as well as Gal and Ghahramani [8] found that MC Dropout, besides providing uncertainty, can also improve the segmentation performance. Since the BraTS challenge evaluation aims at achieving a high segmentation performance, we used the WSCore05 model for the segmentation and survival task at the BraTS17 challenge. Furthermore, Fig. 4 shows that the five models with different Dropout strategies are as well comparable in terms of variation of Dice coefficient and Hausdorff distance (95th percentile). However, in contrast to the average results in Table 1, the median and interquartile range of the MCFull05 model are close to the WSCore05 distribution. This difference can be explained by the rather high amount of outliers in the Dice coefficient variation of the MCFull05 model. Table 2 lists the results achieved on the BraTS17

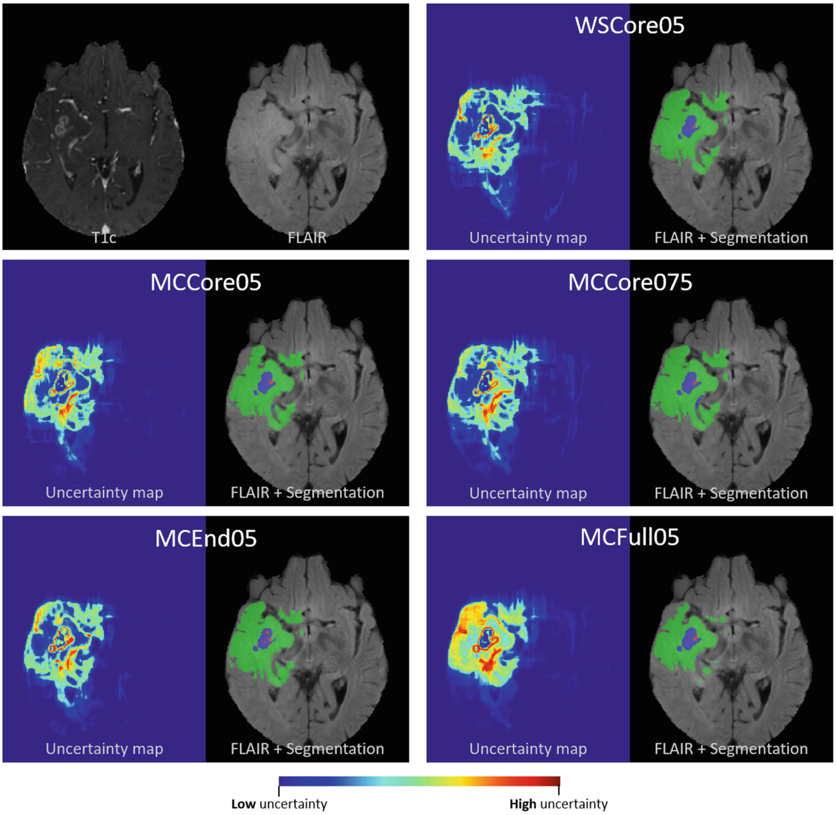


Fig. 5. Uncertainty maps next to FLAIR slices with the segmentation overlaid for each Dropout strategy. The corresponding subject is CBICA_ATW_1 from the BraTS17 validation dataset. As a reference, the raw T1-weighted post-contrast and FLAIR sequences are depicted in the top left.

challenge dataset. Compared to the validation dataset results, the metrics are inferior for all three tumor regions. Reasons might be (a) the performed model selection according to validation dataset results, (b) the difference of the validation dataset and challenge dataset size and (c) a validation dataset that is possibly closer to the training dataset distribution than the challenge dataset.

The qualitative results of the produced uncertainty maps in Fig. 5 highlight that the resulting uncertainty maps of the MC Dropout models MCCore05, MCCore075 and MCEnd05 are visually not distinctively more informative than the entropy determined by the weight scaling model WSCore05 (without MC samples). This problem might come from the rather complex models we use; we could observe that even with a small number of MC samples, the MC Dropout models achieved good segmentation performances (close to the ones shown). It seems that the large complexity allows the models to compensate for the dropped weights and thus minimize the variance in the K MC samples. In contrast to the aforementioned MC Dropout models, the uncertainty produced by the MCFull05 seems to appear more informative. It shows uncertainty where the other models do not (e.g. most parts of the edema), and has increased uncertainty in regions where the other models are uncertain as well. One reason for the more informative uncertainty estimation might be that a fully Bayesian neural network is applied in MCFull05.

The results obtained for the survival prediction on the BraTS validation dataset (Table 3, top row) are not among the best-performing, when comparing it to the results yielding the 2$^{\text{nd}}$ place (Table 3, bottom row) in the challenge. We hypothesize that this difference is due to the rather small validation dataset size ($n = 33$) which might bias the outcomes. Moreover, the likewise small training dataset size could lead to overfitting when using a large number of features. Since our model uses only four features, the chance to overfit is greatly reduced. The avoidance of fine-tuning towards the validation dataset could also play a role in the discrepancy of the results. Furthermore, irregularity of tumor shape turned out to be one of the most predictive features for patient survival which confirms previous findings in literature [6,12,16].

In a next step, we plan to incorporate the generated uncertainty maps into the survival prediction pipeline in order to enhance prediction performance.

5 Conclusion

In conclusion, the results show that the presented models with weight scaling and Monte Carlo Dropout strategies achieve a good segmentation performance and that the visually most informative uncertainty can be obtained by a fully Bayesian neural network (MC Dropout after each convolution layer). First evidence suggests that there might be a trade-off between model complexity and model uncertainty. We could further observe that age and other geometrical features play an important role in survival prediction. Additionally, results in survival prediction indicate a potential prevention of overfitting due to the usage of a small number of features.

Acknowledgments. This work was supported by the Swiss National Foundation by grant number 169607, the Swiss Cancer League project MANAGE KFS-3979-08-2016, the Spanish Ministerio de Economía y Competitividad/FEDER by grant number MTM2015-71200-R, and the James S. Mc. Donnell Foundation 21st Century Science Initiative in Mathematical and Complex Systems Approaches for Brain Cancer [Special Initiative Collaborative – Planning Grant 220020420 and Collaborative award 220020450].

References

1. Alberts, E., Rempfler, M., Alber, G., Huber, T., Kirschke, J., Zimmer, C., Menze, B.H.: Uncertainty quantification in brain tumor segmentation using CRFs and random perturbation models. In: Proceedings - International Symposium on Biomedical Imaging, vol. 2016 June, pp. 428–431. IEEE, April 2016
2. Bakas, S., Akbari, H., Sotiras, A., Bilello, M., Rozycki, M., Kirby, J.S., Freymann, J.B., Farahani, K., Davatzikos, C.: Advancing the cancer genome atlas glioma MRI collections with expert segmentation labels and radiomic features. Nat. Sci. Data **4**, 170117 (2017)
3. Bakas, S., Akbari, H., Sotiras, A., Bilello, M., Rozycki, M., Kirby, J.S., Freymann, J.B., Farahani, K., Davatzikos, C.: Segmentation labels and radiomic features for the pre-operative scans of the TCGA-LGG collection. The Cancer Imaging Archive, January 2017. https://wiki.cancerimagingarchive.net/display/DOI/Segmentation+Labels+and+Radiomic+Features+for+the+Pre-operative+Scans+of+the+TCGA-LGG+collection
4. Bakas, S., Akbari, H., Sotiras, A., Bilello, M., Rozycki, M., Kirby, J.S., Freymann, J.B., Farahani, K., Davatzikos, C.: Segmentation labels and radiomic features for the pre-operative scans of the TCGA-GBM collection. The Cancer Imaging Archive, January 2017. https://wiki.cancerimagingarchive.net/display/DOI/Segmentation+Labels+and+Radiomic+Features+for+the+Pre-operative+Scans+of+the+TCGA-GBM+collection;jsessionid=C2BE9FB8F9D5532DCA9E5CD294787DBC
5. Cui, Y., Tha, K.K., Terasaka, S., Yamaguchi, S., Wang, J., Kudo, K., Xing, L., Shirato, H., Li, R.: Prognostic imaging biomarkers in glioblastoma: development and independent validation on the basis of multiregion and quantitative analysis of MR images. Radiology **278**(2), 546–553 (2016)
6. Czarnek, N., Clark, K., Peters, K.B., Mazurowski, M.A.: Algorithmic three-dimensional analysis of tumor shape in MRI improves prognosis of survival in glioblastoma: a multi-institutional study. J. Neurooncol. **132**(1), 55–62 (2017)
7. Gal, Y.: Uncertainty in Deep Learning. Ph.D. thesis, University of Cambridge (2016)
8. Gal, Y., Ghahramani, Z.: Bayesian convolutional neural networks with Bernoulli approximate variational inference, June 2015. http://arxiv.org/abs/1506.02158
9. Gillies, R.J., Kinahan, P.E., Hricak, H.: Radiomics: images are more than pictures, they are data. Radiology **278**(2), 563–577 (2016)
10. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 770–778 (2016)
11. Kendall, A., Badrinarayanan, V., Cipolla, R.: Bayesian SegNet: model uncertainty in deep convolutional encoder-decoder architectures for scene understanding. In: Proceedings of the British Machine Vision Conference (BMVC) (2017)
12. Kickingereder, P., Neuberger, U., Bonekamp, D., Piechotta, P.L., Götz, M., Wick, A., Sill, M., Kratz, A., Shinohara, R.T., Jones, D.T.W., Radbruch, A., Muschelli, J., Unterberg, A., Debus, J., Schlemmer, H.P., Herold-Mende, C., Pfister, S., von Deimling, A., Wick, W., Capper, D., Maier-Hein, K.H., Bendszus, M.: Radiomic subtyping improves disease stratification beyond key molecular, clinical and standard imaging characteristics in patients with glioblastoma. Neuro-Oncology, nox188 (2017)

13. Lê, M., Unkelbach, J., Ayache, N., Delingette, H.: GPSSI: gaussian process for sampling segmentations of images. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) MICCAI 2015. LNCS, vol. 9351, pp. 38–46. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-24574-4_5
14. Meier, R., Knecht, U., Jungo, A., Wiest, R., Reyes, M.: Perturb-and-MPM: quantifying segmentation uncertainty in dense multi-label CRFs, March 2017. http://arxiv.org/abs/1703.00312
15. Menze, B., Jakab, A., Bauer, S., Kalpathy-Cramer, J., Farahani, K., Kirby, J., Burren, Y., Porz, N., Slotboom, J., Wiest, R., Lanczi, L., Gerstner, E., Weber, M.A., Arbel, T., Avants, B., Ayache, N., Buendia, P., Collins, L., Cordier, N., Corso, J., Criminisi, A., Das, T., Delingette, H., Demiralp, C., Durst, C., Dojat, M., Doyle, S., Festa, J., Forbes, F., Geremia, E., Glocker, B., Golland, P., Guo, X., Hamamci, A., Iftekharuddin, K., Jena, R., John, N., Konukoglu, E., Lashkari, D., Antonio Mariz, J., Meier, R., Pereira, S., Precup, D., Price, S.J., Riklin-Raviv, T., Reza, S., Ryan, M., Schwartz, L., Shin, H.C., Shotton, J., Silva, C., Sousa, N., Subbanna, N., Szekely, G., Taylor, T., Thomas, O., Tustison, N., Unal, G., Vasseur, F., Wintermark, M., Hye Ye, D., Zhao, L., Zhao, B., Zikic, D., Prastawa, M., Reyes, M., Van Leemput, K.: The multimodal brain tumor image segmentation benchmark (BRATS). IEEE Trans. Med. Imaging **34**, 33 (2014)
16. Pérez-Beteta, J., Martínez-González, A., Molina, D., Amo-Salas, M., Luque, B., Arregui, E., Calvo, M., Borrás, J.M., López, C., Claramonte, M., Barcia, J.A., Iglesias, L., Avecillas, J., Albillo, D., Navarro, M., Villanueva, J.M., Paniagua, J.C., Martino, J., Velásquez, C., Asenjo, B., Benavides, M., Herruzo, I., Delgado, M.D.C., del Valle, A., Falkov, A., Schucht, P., Arana, E., Pérez-Romasanta, L., Pérez-García, V.M.: Glioblastoma: does the pre-treatment geometry matter? a postcontrast T1 MRI-based study. Eur. Radiol. **27**(3), 1096–1104 (2017)
17. Pohlen, T., Hermans, A., Mathias, M., Leibe, B.: Full-resolution residual networks for semantic segmentation in street scenes. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2017)
18. Velazquez, R.E., Meier, R., Dunn, W.D., Alexander, B., Wiest, R., Bauer, S., Gutman, D.A., Reyes, M., Aerts, H.J.W.L.: Fully automatic GBM segmentation in the TCGA-GBM dataset: prognosis and correlation with VASARI features. Scientific reports 5, 16822, November 2015
19. Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., Salakhutdinov, R.: Dropout: a simple way to prevent neural networks from overfitting. J. Mach. Learn. Res. **15**, 1929–1958 (2014)
20. Yip, S.S.F., Aerts, H.J.W.L.: Applications and limitations of radiomics. Phy. Med. Biol. **61**(13), R150–R166 (2016)