German Medical Data Sciences 2024
R. Röhrig et al. (Eds.)

doi:10.3233/SHTI240878

# Predicting Overall Survival of Glioblastoma Patients Using Deep Learning Classification Based on MRIs

Katharina OTT[a,1], Santiago CEPEDA[b], Dennis HARTMANN[a], Frank KRAMER[a] and Dominik MÜLLER[a,c]
[a] *IT-Infrastructure for Translational Medical Research, University of Augsburg, Germany*
[b] *Department of Neurosurgery, Río Hortega University Hospital, Spain*
[c] *Institute for Digital Medicine, University Hospital Augsburg, Germany*

**Abstract. Introduction** Glioblastoma (GB) is one of the most aggressive tumors of the brain. Despite intensive treatment, the average overall survival (OS) is 15-18 months. Therefore, it is helpful to be able to assess a patient's OS to tailor treatment more specifically to the course of the disease. Automated analysis of routinely generated MRI sequences (FLAIR, T1, T1CE, and T2) using deep learning-based image classification has the potential to enable accurate OS predictions. **Methods** In this work, a method was developed and evaluated that classifies the OS into three classes – "short", "medium" and "long". For this purpose, the four MRI sequences of a person were corrected using bias-field correction and merged into one image. The pipeline was realized by a bagging model using 5-fold cross-validation and the ResNet50 architecture. **Results** The best model was able to achieve an F1-score of 0.51 and an accuracy of 0.67. In addition, this work enabled a largely clear differentiation of the "short" and "long" classes, which offers high clinical significance as decision support. **Conclusion** Automated analysis of MRI scans using deep learning-based image classification has the potential to enable accurate OS prediction in glioblastomas.

**Keywords.** Glioblastoma, Survival Analysis, Deep Learning

## 1. Introduction

A glioma is caused by the degeneration of glial cells in the brain and can be divided into four grades according to the World Health Organization (WHO) 2016 [1]. Grade I and II gliomas are considered as low-grade gliomas, while grades III and IV are high-grade gliomas. The latter is associated with aggressive growth and a poor prognosis. The most common glioma is the glioblastoma (GB) - grade IV astrocytoma [1]. In 2019-2020, almost 70% of malignant brain tumors in Germany were diagnosed as GB [2]. GB is not only the most diagnosed glioma but also one of the most aggressive and is therefore associated with an unfavorable ("infaust") prognosis. Despite intensive treatment, the average overall survival (OS) is 15-18 months [1].

[1] Dominik Müller, IT Infrastructure for Translational Medical Research, Alter Postweg 101, 86159 Augsburg, Germany; E-mail: dominik.mueller@informatik.uni-augsburg.de

By estimating the OS as accurately as possible, treatment can be better tailored to the patient. It was shown that an estimation by the treating physician often proves to be too subjective and optimistic [3]. Consequently, the most accurate OS prediction possible based on four different MRI sequences (FLAIR, T1, T1CE, and T2) using deep learning classification has great potential.

Various medical image classification models use ensemble learning methods, including bagging [4–6]. Ensemble learning is intended to create more robust and accurate models by combining the predictions of several submodels into one [4]. Bagging calculates a prediction by training submodels, having the same architecture and hyperparameters, using different subsets of the dataset (without the testing dataset). The different models - more precisely their predictions - are combined using pooling functions. In bagging aggregate functions are applied often for this purpose, which weight each model equally and represent rather simple functions (e.g. median) [4,7].

In this work, a method was developed and evaluated to predict the OS of GB patients as accurately as possible based on MRIs using a Convolutional Neural Network (CNN) for classification. For this purpose, a bagging model was implemented, which received 3D images as input containing four MRI sequences of a GB patient.

## 2. Materials and Methods

### *2.1. Dataset*

To be able to train reliable and robust models a sufficient-sized dataset is required. For this reason, our dataset consisted of a combination of four publicly available datasets. These were the RHUH-GBM dataset [8,9], the BraTS Challenge 2020 dataset [10–12], the UCSF-PDGM dataset [9,13,14], and the UPenn-GBM dataset [9,15,16]. The RHUH-GBM dataset was used entirely, whereas the remaining datasets were partially included. The selection criteria for the samples used were defined as follows: The presence of a confirmed diagnosis of glioblastoma, the presence of an age indication, and the availability of OS in days.

Thus, the dataset contained MRIs showing glioblastomas from 1,128 patients. The dataset included images with an image size of either 240x240x155 or 230x230x138 voxels. For each patient, four MRIs from the MRI sequences FLAIR, T1, T1CE, and T2, a segmentation mask of the tumor, age, OS in days, and an annotation label indicating the class of OS were available. The OS classes were created using K-means clustering for three groups. Therefore, three classes were defined for the annotation: "short" ($OS < 170$ days), "medium" ($170 \leq OS \leq 599$ days), and "long" ($OS > 599$ days). The dataset contained 281 short OS cases, 565 medium OS cases, and 282 long OS cases. An exemplary sample of the available images of a patient can be seen in Figure 1.

All images were already coregistrated, resampled and skull stripped. For more detailed information on the dataset demographics, we refer to the excellent paper by Cepeda et al. [17].

### *2.2. Pipeline*

In this work, 3D images were created by combining all four MRI sequences of a person in one image. In the following, these are referred to as multimodality-images.

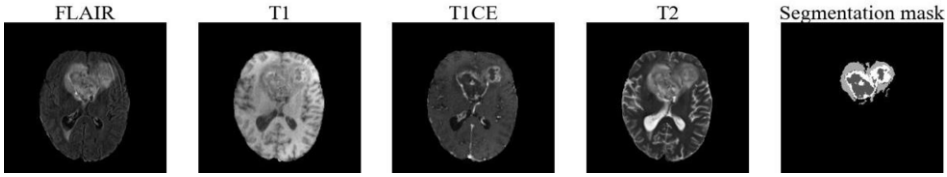


**Figure 1.** Exemplary axial representation of the four MRI sequences in the dataset and the corresponding segmentation mask of the tumor (belonging to one patient).

Before the images were concatenated, preprocessing was performed. This included z-normalization, N4ITK bias-field correction (BFC) [18], resampling, grayscale normalization, padding/cropping, and color format transformation. During the color format transformation, the grayscale values of each voxel were retained by storing the intensity value three times in a tuple. The resulting RGB-format enabled transfer learning for reusing weights fitted on RGB data.

Once the preprocessing was complete, the images were combined. For this purpose, the images per patient were first concatenated along the first axis and then along the second axis. Figure 2 shows an example image of a preprocessed multimodality-image.

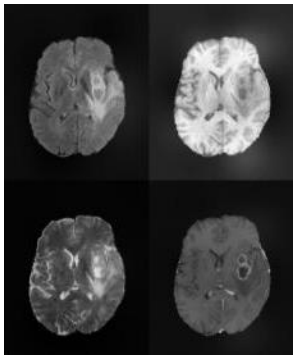

**Figure 2.** Axial representation of a preprocessed 3D multimodality-image.

The images were divided into 80% training data and 20% test data utilizing stratified random sampling in coherence with the recommendations of sampling for machine learning by Sebastian Raschka [19] to avoid overfitting as well as evaluation bias and enable robust model inference. The training dataset was then divided into five subsets using 5-fold cross-validation. In the process, stratified sampling was carried out acknowledging the class distribution. The images were randomly augmented during training using the methods mirroring, rotating, scaling, elastic deforming, changing brightness and contrast to avoid overfitting.

After preprocessing the CNN model received images with a uniform image size of 260x320x103 voxels and three channels. The CNN model was based on a 3D version of the ResNet50 architecture [20,21] already pre-trained on the ImageNet dataset [22]. The architecture was combined with a classification head consisting of a global layer for average pooling, a dense layer with ReLU activation [23], a dropout layer, and a final

dense layer with softmax activation function. For the first 10 epochs, all layers, except those of the classification head, were frozen and trained with a learning rate of $1e^{-4}$. For the fine-tuning, all layers were unfrozen and trained with a dynamic learning rate for a maximum of 240 further epochs, if not aborted by the early-stopping method. The dynamic learning rate started at $1e^{-5}$ and was reduced to a maximum of $1e^{-7}$. A decreasing factor of 0.1 was applied after 8 epochs without improvement of the monitored validation loss. For the Training process, an Adam optimizer [24], weighted focal loss function [25], and a batch size of four were used.

The five homogenous submodels were combined using the following aggregate functions: mean, median, softmax-normalized majority vote soft (MVS), majority vote hard (MVH), and global argmax. MVS normalized the sum of probabilities per class across models using softmax. MVH selected the class with the highest number of votes (vote: highest predicted class) among all models. Global argmax picked the class with the highest probability across all models.

Various metrics were applied to assess the performance of the model per class: accuracy, F1-score, sensitivity, specificity, and AUC. Confusion matrix and ROC curve were also created for the bagging models. For more detailed information on the calculations of the metrics, we refer to the excellent paper by Maier-Hein et al. [26]. The implementation details of our pipeline are summarized in Table 1.

**Table 1.** Overview of configurations for applied preprocessing techniques and neural network models in the presented medical image classification pipeline.

| **Sampling** | | | |
|---|---|---|---|
| **Training dataset** | | | **Testing dataset** |
| 80% + 5-fold cross-validation | | | 20% |
| **Preprocessing** | | | |
| **Image Augmentation** | | | **Subfunctions** |
| Mirror | Random | Standardize | Z-Score normalization |
| Contrast | Factor range: 0.3 to 3<br>Per channel: false | BFC | N4 Bias-field correction-algorithm (shrink factor: 4) |
| Scaling | Factor range: 0.85 to 1.25 | Resampling | $(1.5)^3$ mm |
| Elastic transformation | Alpha: 0 to 900.0<br>Sigma: 9.0 to 13.0 | Padding | To 130x160x103 Voxel (constant factor) |
| Brightness | Factor range: 0.5 to 2<br>Per channel: false | Cropping | To 130x160x103 Voxel (centering) |
| Rotation | Random 90° | Chromer | To RGB format |
| **Neural Network** | | | |
| Loss | Weighted Focal Loss | Class Weights | Computed on train set |
| Batch size | 4 | Optimizer | Adam |
| Learning Rate | Initialized at $1e^{-4}$ (frozen-layer epochs)<br>Initialized at $1e^{-5}$ (unfrozen-layer epochs) | Dynamic learning Rate | Decreasing up to $1e^{-7}$ by a factor of 0.1 each time after 8 epochs without validation loss improvement |
| Epochs | Max 250 | Iterations | #samples/batch_size |
| Transfer Learning | ImageNet | Number of frozen Epochs | 10 |
| Model Checkpoints | Best computed loss, AUC- and F1-score on the validation set during training | Early Stopping | After 32 epochs without validation loss improvement |
| Training Monitoring | CSV dumps for logging | Classification Head | Global average pooling, Dense Layer (ReLU), Dropout Layer and Dense Layer (softmax) |

The bagging models received preprocessed multimodality-images as input and the output consisted of the probability of occurrence of each class, normalized by softmax. The bagging approach was chosen because Müller et al. [4] have shown that it results in a performance and a robustness increase.

All scripts were implemented using the in-house framework AUCMEDI [27], which is based on TensorFlow [28]. The pipeline was executed on a workstation with NVIDIA Titan RTX with 24GB VRAM, Intel(R) Xeon(R) Gold 5220RCPU@2.20GHz with 96 cores and 384GB RAM.

## 3. Results

A deep learning classification model was created, which was implemented as a bagging model using 5-fold cross-validation. The different models were trained with BFC-corrected images on ResNet50. A total of five bagging models were created (combination via five pooling functions) and the pipeline was trained for 331 epochs. As in Table 2 can be seen, the models resulted in similar scores regarding all metrics. All results are represented as the mean over the class-wise results. Therefore, the models obtained an accuracy between 66.5% and 67.7%, an AUC-score of 0.651-0.695, an F1-score of 0.493-0.511, a sensitivity between 51.9% and 53.1% as well as a 74.9%-75.2% specificity. The variance between the classes proved to be very low (<1%) in all metrics except sensitivity (<1.7%). The best result regarding the F1-score was the combination of the five submodels by mean and majority vote soft resulting in 0.511. These models resulted also in the highest scores regarding accuracy (0.677), sensitivity (0.531), and specificity (0.752). The model combined by the MVS classified the "medium" class incorrectly as "long" or "short" in approximately 26% of cases. In under 28% an object of "long" or "short" was assigned to the "medium" class. The classes were correctly classified in 62.5% ("short"), 46.9% ("medium"), and 50% ("long"). In addition, 11% of the images belonging to "short" were assigned to "long". However, 20% of the long-term OS were classified as short-term OS (see Figure 3). More detailed results of the individual models can be found in the GitHub repository linked in the contributions.

**Table 2.** Achieved results of five bagging models showing the Accuracy, AUC, F1-score, Sensitivity, and Specificity on image classification for each pooling function The values shown are the average of the class results. The variance of the classes is shown in brackets behind each value.

| Pooling function | Accuracy | AUC | F1 | Sensitivity | Specificity |
|---|---|---|---|---|---|
| Global argmax | 0.665 (+/-0.0043) | 0.651 (+/- 0.003) | 0.493 (+/- 0.0026) | 0.519 (+/- 0.012) | 0.749 (+/- 0.0000) |
| Majority vote hard | 0.674 (+/- 0.0059) | 0.639 (+/- 0.0038) | 0.505 (+/- 0.0033) | 0.528 (+/- 0.0131) | 0.751 (+/- 0.0013) |
| Majority vote soft | **0.677** (+/- 0.0056) | 0.691 (+/- 0.0044) | **0.511** (+/- 0.001) | **0.531** (+/- 0.0068) | **0.752** (+/- 0.0018) |
| Mean | **0.677** (+/- 0.0056) | 0.690 (+/- 0.0045) | **0.511** (+/- 0.001) | **0.531** (+/- 0.0068) | **0.752** (+/- 0.0018) |
| Median | 0.674 (+/- 0.0057) | **0.695** (+/- 0.0038) | 0.504 (+/- 0.0029) | 0.528 (+/- 0.017) | 0.751 (+/- 0.0027) |

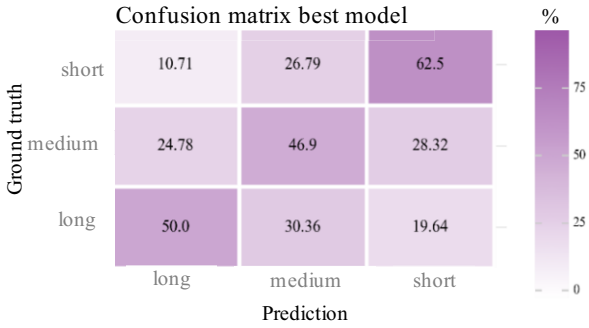


**Figure 3.** Performance results for the best model combined via majority vote soft showing the confusion matrix. A confusion matrix visualizes the comparison of the ground truth values against the predictions of the model.

## 4. Discussion

By combining the submodels by mean/majority vote soft, the best performance in terms of predicting the OS of GB patients was achieved (according to F1-score). A first step towards a clear differentiation of the "short" and "long" classes was achieved, as "short" was only identified as "long" in 10% of cases and vice versa in 20%. The differentiation of the three classes, especially of short-term and long-term survivors is relevant for identifying more specialized treatment. Given the natural course of GB, an accurate prediction of "short" may indicate that treatment should be withheld to prioritize quality of life over therapeutic obstinacy. Conversely, intensive therapy could be considered in correctly predicted long-term survivors. Nevertheless, a prediction of OS in days should follow and needs to be further investigated in the future.

The performance of all models could be addressed to bagging. In bagging, homogeneous models are combined by pooling functions so that the prediction of the ensemble model proves to be more robust and accurate [4]. The five submodels in this work were each trained on different subsets of the dataset. This allowed individual models to learn slightly different features as different images were available for learning. Thus, one model may have been able to recognize a feature through several very extreme cases of short-term OS that another submodel could not learn because there were not as many extreme cases in the training set. A limiting factor was the restriction of pooling functions to aggregate functions. The use and impact of other, more complex pooling functions such as Random Forest or Support Vector Machine should be investigated.

Some studies show that a combination of different MRI sequences, in one image, provides more features than a single sequence, favoring a more accurate classification [29,30]. Our bagging models were able to achieve high scores, especially in the metrics AUC, accuracy, and specificity, which supports the assumption of providing additional features. Liang et al. [29] and Coupet et al. [30] showed in their work that combining different MRI sequences along the channel axis shows great potential. A combination in this way should be compared with the method described here and investigated further, which wouldn't increase the voxel resolution of two axes, but of one axis.

A comparison of our model with state-of-the-art models used for predicting OS of GB patients can be seen in Table 3. We have limited ourselves to the best-placed papers in the BraTS Challenge 2020, as they have a similar dataset. Our best model achieved an accuracy of 0.677. This corresponds to a higher value than the winners of the BraTS

Challenge 2020 McKinley et al. [31] and Asemio and Solís [32], who achieved 0.617. As shown in Table 3, all papers use different approaches, but the most popular methods consist of z-score normalization, an ensemble approach, and the use of radiomic/image-based features. We have additionally used multimodality images (possibly adding more features), CNNs (popular for image-based deep learning), transfer learning, and BFC (improved performance of CNNs). Nevertheless, our model has some limitations and weaknesses in terms of methods. The models in the comparison (Table 3) and Baid et al. [33] often use radiomic features. Especially age has been shown to have a high correlation with OS. Similarly, segmented images are often used to provide a more detailed view of the tumor. Both the inclusion of age and the use of segmented images should be further investigated in the future. In addition, regression is always used to predict the exact OS in days. An ensemble of our approach with a regression should be investigated for more accurate results. However, a more detailed comparison with their methodologies on the same dataset should be further investigated for better comparability.

**Table 3.** Comparison in terms of the dataset(s), methods, and metrics of our model with the models of the top-ranked participants of the BraTS Challenge 2020, which are also used to predict the OS of GB patients.

| Literature | Dataset(s) | Methods | Metrics/Results |
|---|---|---|---|
| McKinley et al. [31] | BraTS 2020 [10–12] | Normalization/homogenization: z-score<br>ordinary least square regression model (OLS) and Random Forest (RF) classification model<br>Using age, number of tumor cores, and number of tumor regions<br>Ensemble of OLS and RF model | Accuracy: 0.617 |
| Asemjo and Solís [32] | BraTS 2020 [10–12] | Segmented images<br>3 models: classification (Decision tree and SVM) and regression model (Ensemble regression trees)<br>Using image-based/radiomic features and age<br>Ensemble of the three models (mean) | Accuracy: 0.617 |
| Bommineni [34] | BraTS 2020 [10–12] | Normalization: z-score<br>Segmented images (tumorous regions)<br>Random Forest regressor<br>Using image-based/radiomic features and age | Accuracy: 0.589 |
| Ali et al. [35] | BraTS 2020 [10–12] | Normalization: z-score<br>Linear Regression model<br>Using age, surface area, volume, spatial location, and resection status | Accuracy: 0.579 |
| Our model | BraTS 2020, UCSF-PDGM, UPenn-GBM, RHUH [8–16] | Normalization: z-score<br>BFC<br>Multimodality-images<br>Transfer Learning<br>CNN model<br>Using MRIs, image-based features<br>Ensemble learning (Bagging) | Accuracy: 0.677 |

## 5. Conclusion

In this study, deep learning classification models were developed and evaluated to predict the OS of GB patients based on 3D images consisting of four MRI sequences. These images were used as input to a CNN model that utilized ensemble learning and the

ResNet50 architecture. Different pooling functions were tested. It could be shown that the two classes "short" and "long" can be largely differentiated from each other, which could enable individualized therapy. Furthermore, relatively accurate and robust models for the prediction of OS of GB patients could be generated using deep learning. However, these could be further improved by including radiomics features or by integrating the age of the patients. The influence of more complex pooling functions on the bagging model and the ensemble with a regression should also be analyzed.

## Declarations

### *Code Availability*

The code for this article was implemented in Python and is available under the GPL-3.0 License at the following GitHub repository: https://github.com/frankkramer-lab/Glioblastoma-Survival-Prognosis-via-Image-Classification.git.

### *Competing Interests*

The authors declare no conflicts of interest.

### *Author Contributions*

F.K. contributed to the coordination and funding of this work. S.C. provided clinical expertise. D.H. participated in reviewing the manuscript. D.M. was responsible for supervision and contribution to the manuscript. K.O. was in charge of drafting the manuscript and implementing the code. All authors have read and agreed to the published version of the manuscript.

### *Ethics Approval and Consent to Participate*

Not applicable.

## References

1. Gritsch S, Batchelor TT, Gonzalez Castro LN. Diagnostic, therapeutic, and prognostic implications of the 2021 World Health Organization classification of tumors of the central nervous system. Cancer. 2022 Jan;128(1):47–58, doi: 10.1002/cncr.33918
2. Robert Koch-Institut, Gesellschaft der epidemiologischen Krebsregister in Deutschland e.V., editors. Krebs in Deutschland für 2019/2020. 14th ed. Berlin: Robert Koch-Institut; 2023. 160 p, doi: 10.25646/11357
3. Glare P, Virik K, Jones M, Hudson M, Eychmuller S, Simes J, Christakis N. A systematic review of physicians' survival predictions in terminally ill cancer patients. BMJ. 2003 Jul;327(7408):195–8, doi: 10.1136/bmj.327.7408.195
4. Müller D, Soto-Rey I, Kramer F. An Analysis on Ensemble Learning optimized Medical Image Classification with Deep Convolutional Neural Networks. IEEE Access. 2022 Jun;10:66467-80, doi: 10.1109/ACCESS.2022.3182399
5. Hameed Z, Zahia S, Garcia-Zapirain B, Javier Aguirre J, María Vanegas A. Breast cancer histopathology image classification using an ensemble of deep learning models. Sensors. 2020 Aug;20(16):4373, doi: 103390/s20164373

6. Liu Y, Long F. Acute lymphoblastic leukemia cells image analysis with deep bagging ensemble learning. In: Gupta A, Gupta R, editors. ISBI 2019 C-NMC challenge: Classification in cancer cell imaging. Singapore: Springer Singapore; c2019. p. 113–21, doi: 10.1007/978-981-15-0798-4_12
7. Müller D. Frameworks in medical image analysis with deep neural networks [doctoralthesis]. Augsburg: Universität Augsburg; 2023. 317 p
8. Cepeda S, García-García S, Arrese I, Herrero F, Escudero T, Zamora T, Sarabia R. The río hortega university hospital glioblastoma dataset: A comprehensive collection of preoperative, early postoperative and recurrence MRI scans (RHUH-GBM). Data Brief. The Cancer Imaging Archive. 2023, doi: 10.7937/4545-c905
9. Clark K, Vendt B, Smith K, Freymann J, Kirby J, Koppel P, Moore S, Phillips S, Maffitt D, Pringle M, Tarbox L, Prior F. The Cancer Imaging Archive (TCIA): Maintaining and Operating a Public Information Repository. J Digit Imaging. 2013 Dec;26(6):1045–57, doi: 10.1007/s10278-013-9622-7
10. Menze BH, Jakab A, Bauer S, Kalpathy-Cramer J, Farahani K, Kirby J, Burren Y, Porz N, Slotboom J, Wiest R, Lanczi L, Gerstner E, Weber MA, Arbel T, Avants BB, Ayache N, Buendia P, Collins DL, Cordier N, Corso JJ, Criminisi A, Das T, Delingette H, Demiralp Ç, Durst CR, Dojat M, Doyle S, Festa J, Forbes F, Geremia E, Glocker B, Golland P, Guo X, Hamamci A, Iftekharuddin KM, Jena R, John NM, Konukoglu E, Lashkari D, Mariz JA, Meier R, Pereira S, Precup D, Price SJ, Raviv TR, Reza SMS, Ryan M, Sarikaya D, Schwartz L, Shin HC, Shotton J, Silva CA, Sousa N, Subbanna NK, Szekely G, Taylor TJ, Thomas OM, Tustison NJ, Unal G, Vasseur F, Wintermark M, Ye DH, 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. 2015 Oct;34(10):1993–2024, doi: 10.1109/TMI.2014.2377694
11. Bakas S, Reyes M, Jakab A, Bauer S, Rempfler M, Crimi A, Shinohara R, Berger C, Ha S, Rozycki M, Prastawa M, Alberts E, Lipkova J, Freymann J, Kirby J, Bilello M, Fathallah-Shaykh H, Wiest R, Kirschke J, Chen Z. Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the BRATS challenge. 2019 Mar;38, doi: 10.48550/arXiv.1811.02629
12. Bakas S, Akbari H, Sotiras A, Bilello M, Rozycki M, Kirby JS, Freymann JB, Farahani K, Davatzikos C. Advancing The Cancer Genome Atlas glioma MRI collections with expert segmentation labels and radiomic features. Sci Data. 2017 Sep;4:170117, doi: 10.1038/sdata.2017.117
13. Calabrese E, Villanueva-Meyer JE, Rudie JD, Rauschenecker AM, Baid U, Bakas S, Cha S, Mongan JT, Hess CP. The University of California San Francisco Preoperative Diffuse Glioma MRI (UCSF-PDGM). Data Brief. The Cancer Imaging Archive. 2022;4, doi: 10.7937/tcia.bdgf-8v37
14. Calabrese E, Villanueva-Meyer JE, Rudie JD, Rauschecker AM, Baid U, Bakas S, Cha S, Mongan JT, Hess CP. The university of california san francisco preoperative diffuse glioma MRI dataset. Radiol Artif Intell. 2022 Oct;4(6):e220058, doi: 10.1148/ryai.220058
15. Bakas S, Sako C, Akbari H, Bilello M, Sotiras A, Shukla G, Rudie JD, Santamaría NF, Kazerooni AF, Pati S, Rathore S, Mamourian E, Ha SM, Parker W, Doshi J, Baid U, Bergman M, Binder ZA, Verma R, Lustig RA, Desai AS, Bagley SJ, Mourelatos Z, Morrissette J, Watt CD, Brem S, Wolf RL, Melhem ER, Nasrallah MP, Mohan S, O'Rourke DM, Davatzikos C. Multi-parametric magnetic resonance imaging (mpMRI) scans for de novo Glioblastoma (GBM) patients from the University of Pennsylvania Health System (UPENN-GBM). Data Brief. The Cancer Imaging Archive. 2021;2, doi: 10.7937/TCIA.709X-DN49
16. Bakas S, Sako C, Akbari H, Bilello M, Sotiras A, Shukla G, Rudie JD, Santamaría NF, Kazerooni AF, Pati S, Rathore S, Mamourian E, Ha SM, Parker W, Doshi J, Baid U, Bergman M, Binder ZA, Verma R, Lustig RA, Desai AS, Bagley SJ, Mourelatos Z, Morrissette J, Watt CD, Brem S, Wolf RL, Melhem ER, Nasrallah MP, Mohan S, O'Rourke DM, Davatzikos C. The University of Pennsylvania glioblastoma (UPenn-GBM) cohort: advanced MRI, clinical, genomics, & radiomics. Sci Data. 2022 Jul;9(1):453, doi: 10.1038/s41597-022-01560-7
17. Cepeda S, García-García S, Arrese I, Herrero F, Escudero T, Zamora T, Sarabia R. The Río Hortega University Hospital Glioblastoma dataset: A comprehensive collection of preoperative, early postoperative and recurrence MRI scans (RHUH-GBM). Data Brief. 2023 Oct;50:109617, doi: 10.1016/j.dib.2023.109617
18. Tustison NJ, Avants BB, Cook PA, Zheng Y, Egan A, Yushkevich PA, Gee JC. N4ITK: improved N3 bias correction. IEEE Trans Med Imaging. 2010 Jun;29(6):1310–20, doi: 10.1109/TMI.2010.2046908
19. Raschka S. Model evaluation, model selection, and algorithm selection in machine learning. ArXiv. 2018 Nov, doi: 10.48550/arXiv.1811.12808
20. Solovyev R, Kalinin AA, Gabruseva T. 3D convolutional neural networks for stalled brain capillary detection. Comput Biol Med. 2022 Feb;141:105089, doi: 10.1016/j.compbiomed.2021.105089
21. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2016:770-8, doi: 10.1109/CVPR.2016.90

22. Russakovsky O, Deng J, Su H, Krause J, Satheesh S, Ma S, Huang Z, Karpathy A, Khosla A, Bernstein M, Berg AC, Fei-Fei L. ImageNet Large Scale Visual Recognition Challenge. Int J Comput Vis. 2015 Dec;115(3):211–52, doi: 10.1007/s11263-015-0816-y
23. Agarap AF. Deep learning using rectified linear units (ReLU). ArXiv. 2018 Mar;abs/1803.08375, doi: 10.48550/arXiv.1803.08375
24. Kingma DP, Ba J. Adam: A method for stochastic optimization. ArXiv. 2017 Jan;9, doi: 10.48550/arXiv.1412.6980
25. Lin TY, Goyal P, Girshick R, He K, Dollár P. Focal loss for dense object detection. IEEE Trans Pattern Anal Mach Intell. 2020 Feb;42(2):318–27, doi: 10.1109/TPAMI.2018.2858826
26. Maier-Hein L, Reinke A, Godau P, Tizabi MD, Buettner F, Christodoulou E, Glocker B, Isensee F, Kleesiek J, Kozubek M, Reyes M, Riegler MA, Wiesenfarth M, Kavur AE, Sudre CH, Baumgartner M, Eisenmann M, Heckmann-Nötzel D, Rädsch T, Acion L, Antonelli M, Arbel T, Bakas S, Benis A, Blaschko MB, Cardoso MJ, Cheplygina V, Cimini BA, Collins GS, Farahani K, Ferrer L, Galdran A, van Ginneken B, Haase R, Hashimoto DA, Hoffman MM, Huisman M, Jannin P, Kahn CE, Kainmueller D, Kainz B, Karargyris A, Karthikesalingam A, Kofler F, Kopp-Schneider A, Kreshuk A, Kurc T, Landman BA, Litjens G, Madani A, Maier-Hein K, Martel AL, Mattson P, Meijering E, Menze B, Moons KGM, Müller H, Nichyporuk B, Nickel F, Petersen J, Rajpoot N, Rieke N, Saez-Rodriguez J, Sánchez CI, Shetty S, van Smeden M, Summers RM, Taha AA, Tiulpin A, Tsaftaris SA, Van Calster B, Varoquaux G, Jäger PF. Metrics reloaded: recommendations for image analysis validation. Nat Methods. 2024 Feb;21(2):195–212, doi: 0.1038/s41592-023-02151-z
27. Müller D, Hartmann D, Soto-Rey I, Kramer F. Abstract: AUCMEDI. In: Deserno TM, Handels H, Maier A, Maier-Hein K, Palm C, Tolxdorff T, editors. Bildverarbeitung für die medizin 2023; Wiesbaden: Springer Fachmedien Wiesbaden; c2023. p. 253, doi: 10.1007/978-3-658-41657-7_55
28. Abadi M, Agarwal A, Barham P, Brevdo E, Chen Z, Citro C, Corrado GS, Davis A, Dean J, Devin M, Ghemawat S, Goodfellow I, Harp A, Irving G, Isard M, Jia Y, Jozefowicz R, Kaiser L, Kudlur M, Levenberg J, Mane D, Monga R, Moore S, Murray D, Olah C, Schuster M, Shlens J, Steiner B, Sutskever I, Talwar K, Tucker P, Vanhoucke V, Vasudevan V, Viegas F, Vinyals O, Warden P, Wattenberg M, Wicke M, Yu Y, Zheng X. TensorFlow: Large-scale machine learning on heterogeneous distributed systems. ArXiv. 2016 Mar;2, doi: 10.48550/arXiv.1603.04467
29. Liang S, Zhang R, Liang D, Song T, Ai T, Xia C, Xia L, Wang Y. Multimodal 3D DenseNet for IDH genotype prediction in gliomas. Genes. 2018 Jul;9(8):382, doi: 10.3390/genes9080382
30. Coupet M, Urruty T, Leelanupab T, Naudin M, Bourdon P, Maloigne CF, Guillevin R. A multi-sequences MRI deep framework study applied to glioma classfication. Multimed Tools Appl. 2022 Feb;81(10):13563–91, doi: 10.1007/s11042-022-12316-1
31. McKinley R, Rebsamen M, Dätwyler K, Meier R, Radojewski P, Wiest R. Uncertainty-driven refinement of tumor-core segmentation using 3D-to-2D networks with label uncertainty. In: Crimi A, Bakas S, editors. Brainlesion: Glioma, multiple sclerosis, stroke and traumatic brain injuries. Cham: Springer International Publishing; 2021. p. 401–11, doi: 10.1007/978-3-030-72084-1_36
32. Marti Asenjo J, Martinez-Larraz Solís A. MRI brain tumor segmentation using a 2D-3D u-net ensemble. In: Crimi A, Bakas S, editors. Brainlesion: Glioma, multiple sclerosis, stroke and traumatic brain injuries. Cham: Springer International Publishing; 2021. p. 354–66, doi: 10.1007/978-3-030-72084-1_32
33. Baid U, Rane SU, Talbar S, Gupta S, Thakur MH, Moiyadi A, Mahajan A. Overall Survival Prediction in Glioblastoma With Radiomic Features Using Machine Learning. Front Comput Neurosci. 2020 Aug;14:61, doi: 10.3389/fncom.2020.00061
34. Bommineni VL. PieceNet: a redundant unet ensemble. In: Crimi A, Bakas S, editors. Brainlesion: Glioma, multiple sclerosis, stroke and traumatic brain injuries. Cham: Springer International Publishing; 2021. p. 331–41, doi: 10.1007/978-3-030-72087-2_29
35. Ali MJ, Akram MT, Saleem H, Raza B, Shahid AR. Glioma segmentation using ensemble of 2D/3D u-nets and survival prediction using multiple features fusion. In: Crimi A, Bakas S, editors. Brainlesion: Glioma, multiple sclerosis, stroke and traumatic brain injuries. Cham: Springer International Publishing; 2021. p. 189–99, doi: 10.1007/978-3-030-72087-2_17