

Article

Ensembles of Convolutional Neural Networks for Survival Time Estimation of High-Grade Glioma Patients from Multimodal MRI

Kaoutar Ben Ahmed ^{1,*} , Lawrence O. Hall ¹, Dmitry B. Goldgof ¹ and Robert Gatenby ²

¹ Department of Computer Science and Engineering, University of South Florida, Tampa, FL 33620, USA; lohall@usf.edu (L.O.H.); goldgof@usf.edu (D.B.G.)

² Department of Diagnostic Imaging and Interventional Radiology, H. Lee Moffitt Cancer Center and Research Institute, Tampa, FL 33612, USA; Robert.Gatenby@moffitt.org

* Correspondence: kbenahmed@usf.edu

Abstract: Glioma is the most common type of primary malignant brain tumor. Accurate survival time prediction for glioma patients may positively impact treatment planning. In this paper, we develop an automatic survival time prediction tool for glioblastoma patients along with an effective solution to the limited availability of annotated medical imaging datasets. Ensembles of snapshots of three dimensional (3D) deep convolutional neural networks (CNN) are applied to Magnetic Resonance Image (MRI) data to predict survival time of high-grade glioma patients. Additionally, multi-sequence MRI images were used to enhance survival prediction performance. A novel way to leverage the potential of ensembles to overcome the limitation of labeled medical image availability is shown. This new classification method separates glioblastoma patients into long- and short-term survivors. The BraTS (Brain Tumor Image Segmentation) 2019 training dataset was used in this work. Each patient case consisted of three MRI sequences (T1CE, T2, and FLAIR). Our training set contained 163 cases while the test set included 46 cases. The best known prediction accuracy of 74% for this type of problem was achieved on the unseen test set.

Keywords: glioblastoma multiforme; brain tumor; survival prediction; magnetic resonance images; BraTS; artificial intelligence; machine learning; deep learning



Citation: Ben Ahmed, K.; Hall, L.O.; Goldgof, D.B.; Gatenby, R. Ensembles of Convolutional Neural Networks for Survival Time Estimation of High-Grade Glioma Patients from Multimodal MRI. *Diagnostics* **2022**, *12*, 345. <https://doi.org/10.3390/diagnostics12020345>

Academic Editor: Alessio Imperiale

Received: 15 December 2021

Accepted: 24 January 2022

Published: 29 January 2022

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Copyright: © 2022 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

A glioma is the most common primary tumor of the brain. It originates from glial cells. About 80% of malignant brain tumors are gliomas. According to the 2021 WHO Classification of Tumors of the Central Nervous System [1], glioblastoma is a sub-category of adult-type diffuse gliomas. It is considered a WHO grade IV glioma which represents approximately 57% of all gliomas and 48% of all primary malignant central nervous system (CNS) tumors [2]. Glioblastoma is highly invasive and can quickly spread throughout every region of the brain. Standard therapy consists of resection (surgery) followed by a combination of radiation and chemotherapy. Despite aggressive multimodal treatment, patients diagnosed with glioblastoma multiforme (GBM) have a dismal prognosis with an average survival time of slightly more than 1 year (15–16 months) [3].

Neuro imaging remains the cornerstone of tumor diagnosis and assessment during therapy. While treatment response continues to simply be assessed on changes in tumor size, new image analytic tools have been introduced to access additional information from clinical scans. Radiomics is a non-invasive method for the extraction of quantitative features from medical imaging that may not be apparent through traditional visual inspection [4]. By far, the most common neuro-imaging tool is magnetic resonance imaging (MRI). MRI provides an isotropic three dimensional (3D) picture of the brain with excellent soft tissue contrast and resolution without potentially harmful radiation. Typically, MRI imaging is

performed in three different planes; axial, coronal, and sagittal. Figure 1 shows a slice of the human brain in three planes. Commonly used MRI sequences include contrast enhanced T1-weighted (T1CE), T2-weighted, and fluid attenuation inversion recovery (FLAIR) as shown in Figure 2. These sequences, because they are sensitive to different components of the tumor biology, play an integral role in providing refined anatomic and physiological detail. For instance, T1-weighted post-contrast images are sensitive to the enhancing core and necrotic tumor core, whereas FLAIR and T2 highlight the peritumoral edema.

Recently, deep learning methods have been applied in computer vision and biomedical image analysis to improve the ability to extract features from images automatically and increase predictive power. They rely on advances in neural network architectures, GPU computing, and the advent of open source frameworks such as Tensorflow [5] and Pytorch [6]. When trained with sufficient data, deep neural networks can attain high accuracy in medical image analysis. The present state of the art is largely dominated by convolutional neural networks (CNN) with a number of effective architectures such as VGG-Net [7], Inception networks [8], and ResNet [9]. Instead of manually engineering features, convolutional neural networks allow automatic convolutions to extract an enormous number of features from input images. Recent review works [10] show the domination of convolutional neural network techniques applied to brain magnetic resonance imaging (MRI) analysis compared to other approaches.

Survival time after a patient is diagnosed with high grade glioma depends on several factors such as age, treatment, tumor size, behavior, and location, as well as histologic and genetic markers [11]. Integration of automated image analytics can provide insights for diagnosis, therapy planning, monitoring, and prognosis prediction of gliomas. A new study [12] by UT Southwestern shows deep learning models can automatically classify IDH mutation status in brain gliomas from 3D MR imaging with more than 97% accuracy. This technology can potentially eliminate the current need for brain cancer patients to undergo invasive surgical procedures to help doctors determine the best treatment for their tumors. This represents an important milestone towards non-surgical strategies for histological determination especially in highly eloquent brain areas such as the brain stem.

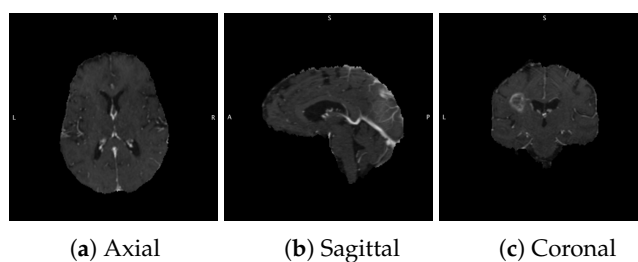


Figure 1. Different planes of brain MRI with the T1CE sequence.

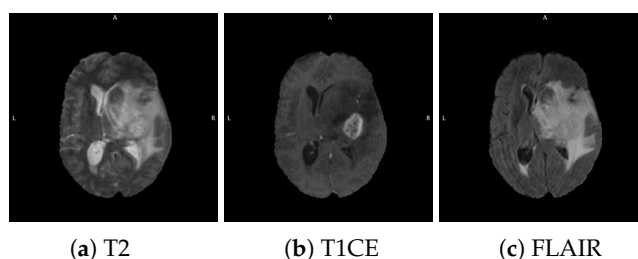


Figure 2. Different sequences of brain MRI in the axial plane.

Applying deep learning algorithms to multi-sequence MR images is challenging due to several factors. First, the size of the 3D data. Second, the difficulty in designing an appropriate neural network architecture for the desired objective. Third (and most critically), the limited clinical knowledge of deep learning experts and the limited availability of large

medical image datasets with labels. Finally, predictive power of deep convolutional neural network models is significantly enhanced by a large training dataset. Using a small number of cases to train a model using deep learning can result in over-fitting the training data such that generalization to future cases is poor.

Here, we address the issues of limited and heterogeneous data sets in medical imaging using an ensemble learning method applied to survival time prediction from glioblastoma patient data. Its efficacy is demonstrated through application to MR images from the BraTS dataset.

We present our investigation as follows: Section II discusses related research studies and recent trends of survival analysis and radiomics using deep learning. In Section III, we describe our proposed survival prediction system in detail including data acquisition, preparation and augmentation, region of interest segmentation and model construction, and training. Our experimental results are presented in Section V. In Section VI, the results are discussed and conclusions are drawn from experimental outcomes.

2. Related Work

The task of survival prediction of glioma from MRI images is challenging. Several studies applying various methods and approaches using the BraTS dataset are reviewed in this section.

Authors of [13–15] proposed the use of some handcrafted and radiomics features extracted from automatically segmented volumes with region labels to train a random forest regression model to predict the survival time of GBM patients in days. These studies achieved respectively 52%, 51.7%, and 27.25% accuracy on the validation set of the BraTS (Brain Tumor Image Segmentation) 2019 challenge.

Authors of [16] performed a two category (short- and long-term) survival classification task using a linear discriminant classifier that was trained with deep features extracted from a pre-trained convolutional neural network (CNN). The study achieved 68.8% accuracy when doing 5-fold cross validation on the BraTS 2017 dataset.

Ensemble learning was used by authors in [17]. They extracted handcrafted and radiomics features from automatically segmented MRI images of high grade gliomas and created an ensemble of multiple classifiers, including random forests, support vector machines, and multilayer perceptrons, to predict overall survival (OS) time on the BraTS 2018 testing set. They obtained an accuracy of 52%.

The authors of [18] achieved first place in the BraTS 2020 challenge for the overall survival prediction task (61.7% accuracy). They extracted segmentation features along with patient's age to classify the patients into three groups (long-, short- and, mid-term survivors) using an ensemble of a linear regression model and random forests classifier.

Over the last decade, there was increasing interest in ensemble learning for tumor segmentation tasks as well. Ensemble learning was ubiquitous in the BraTS 2017–2020 challenges, being used in almost all of the top-ranked methods. The winner of the BraTS 2017 challenge for GBM tumor segmentation [19] was an ensemble of two fully convolutional network models (FCN), and a U-net each generating separate class confidence maps. Then, each class was created by averaging the confidence maps of the individual ensemble models for each voxel. This study reached dice scores of 0.90, 0.82, and 0.75 for the whole tumor, tumor core, and enhancing tumor, respectively for the BraTS 2017 validation set. Authors in [20] built an ensemble of UNet-based deep networks trained in a multi-fold setting to perform segmentation of brain tumors from the T2 Fluid Attenuated Inversion Recovery (T2-FLAIR) sequences. They achieved a dice score of 0.882 for the BraTS 2018 set.

3. Materials and Methods

3.1. Data Acquisition and Preparation

3.1.1. BraTS Dataset

The Brain Tumor Segmentation (BraTS) Challenge is critical for benchmarking and has largely contributed to advancing machine learning applications in glioma image analysis.

The BraTS challenge was first held in 2012 and has taken place annually as part of the Medical Image Computing and Computer Assisted Intervention (MICCAI) conference ever since. It focuses on evaluating state-of-the-art methods for the segmentation of brain tumors in multimodal magnetic resonance imaging (MRI) scans. The glioma sub-regions considered for segmentation evaluation are: (1) The enhancing tumor (ET), (2) the tumor core (TC), and (3) the whole tumor (WT). The enhancing tumor sub-region is described by areas that are typically hypo-intense in T1CE when compared to T1. The tumor core (TC) consists of the bulk of the tumor, which is what is typically resected. The TC entails the ET, as well as the necrotic (fluid-filled) and the non-enhancing (solid) parts of the tumor. The WT corresponds to the complete extent of the disease, as it consists of the TC and the peritumoral edema (ED) [21].

To evaluate proposed segmentation methods, the participants are asked to upload their segmentation labels for regions in the image (e.g., edema, tumor) as a single multi-label file in the NIfTI format. Then metrics like dice score and the Hausdorff distance are reported to rank teams performances. In 2017, the BraTS challenge started to include another task, the prediction of overall survival time. Participants needed to use their produced segmentation in combination with other features to attempt to predict patient overall survival. The evaluation assessment of this task was based on the accuracy metric of a three category classification (long-survivors (e.g., >15 months), short-survivors (e.g., <10 months), and mid-survivors (e.g., between 10 and 15 months)).

The Brain Tumor Segmentation (BraTS) challenge dataset is the largest publicly available glioma imaging dataset. It includes multi-institutional pre-operative clinically-acquired multi-sequence MRI scans of glioblastoma (GBM/HGG) and lower grade glioma (LGG) with overall survival data. Images were acquired from different institutions using MR scanners with different field strengths (1.5 T and 3 T). Segmentation ground truth labels are done by expert board-certified neuroradiologists. Scans are provided in NIfTI format and have (a) native (T1) and (b) post-contrast T1-weighted (T1CE), (c) T2-weighted (T2), and (d) T2 Fluid Attenuated Inversion Recovery (T2-FLAIR) sequences with 1–6-mm slice thickness. The BraTS data was made available after its pre-processing, i.e., co-registered to the same anatomical template, interpolated to the same resolution, and skull-stripped. The overall survival data is defined in days.

3.1.2. Data Acquisition

In this study we used subsets of glioblastoma (GBM) cases from the BraTS 2019 training dataset. We randomly split the data into train/test subsets. The training subset consisting of 163 cases and the testing set had 46 cases. Among the 163 cases, there were 81 long-term and 82 short-term survival cases based on a 12 month cut-off (which is the middle threshold of the mid-term survivors class in BraTS data). Among the 46 test cases, there were 23 long-term and 23 short-term survivors. The data can be downloaded online from the CBICA's Image Processing Portal (IPP) after requesting permission from the BraTS team. Figure 3 shows the Kaplan–Meier survival graph corresponding to the training dataset.

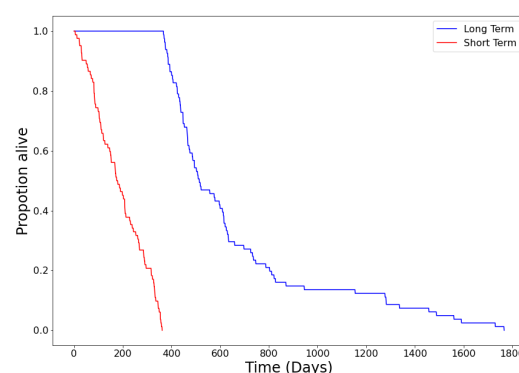


Figure 3. Kaplan–Meier survival graph of the training set.

3.1.3. Data Pre-Processing and Augmentation

The first step is tumor segmentation, we decided to use the publicly available pre-trained model of Wang et al. [22] to automatically segment the training and testing set used in this study. The segmented volumes were used as a starting point for our task of survival class prediction.

Each MRI volume is high resolution with a size of $240 \times 240 \times 155$ which makes it difficult to fit into the GPU memory. To solve this issue, we had to remove empty voxels and null slices that are unnecessary for building a good predictive model.

Data augmentation is a fundamental way to help reduce model over-fitting when dealing with a small amount of training data. Hence, we performed transformations to each 2D slice consisting of vertical flipping and rotations of 10 degrees incrementally from 0 to 180 degrees. In addition, we applied elastic deformation by generating smooth deformations using random displacement fields that are sampled from a Gaussian distribution with a standard deviation of 17 (in pixels). The new values for each pixel were then calculated using bicubic interpolation. The augmentation was applied to each 2D slice which then gets stacked back in to form a 3D cube. The transformations resulted in increasing the size of training data from 163 images per sequence to roughly 10,000 augmented images, including the original images. Then all the images were resized to a uniform size of $160 \times 160 \times 110$. Furthermore, we used the Keras [23] data generator to feed data by batches to our model instead of loading all data into memory at once. In addition, before feeding input images to the model, as a part of preprocessing, the images were normalized to have zero mean and unit standard deviation.

3.2. Overall Survival Prediction System

Our survival prediction system is next described in detail including data acquisition, preparation and augmentation, region of interest segmentation, model construction, and training.

3.2.1. System Outline

As shown in Figure 4, the overall work flow of our system consists of three stages. Firstly, the starting point was the raw 3D MRI images acquired from the BraTS platform after securing permission. The images were sent as inputs to the pre-trained segmentation network [22] resulting in a 3D segmented tumor in the NIfTI format. Image augmentation and pre-processing were done before the inputs were sent to our 3D convolutional neural network for automatic feature extraction and classifier training. Finally, the end point is the overall survival prediction. Ensemble learning was used to evaluate the prediction performance of the trained networks. More details on each phase are provided in the following sub-sections.

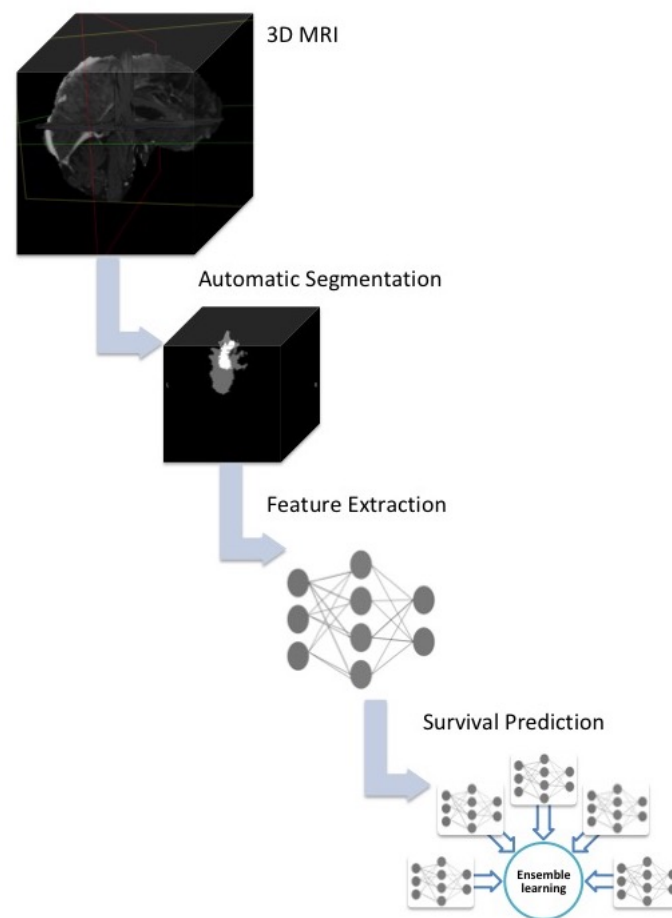


Figure 4. Pipeline of proposed overall survival prediction system.

3.2.2. Snapshot Learning for Survival Prediction

An ensemble of deep neural networks is known to be more accurate and robust. However, training multiple deep networks is computationally expensive. Huang et al. [24] introduced snapshot learning which is a fast and effective way of creating an ensemble of models without additional training costs by saving one model's weights snapshots at different points over the course of training. Our preliminary results in [25] showed the effectiveness of an ensemble of snapshots in boosting the prediction performance using our limited size dataset. Snapshots of the model are saved at every training iteration then a selected subset of them are used to evaluate the model's prediction performance on a separate testing set. Majority voting was used to combine their predictions. The snapshots of the model are each evaluated and a classification outcome is obtained for each of the 46 test cases. Each sample can be correctly or incorrectly classified by a particular snapshot model. The predictions each provide a vote and our combination approach is to take the class with the most votes for each test example. The voting rule is as follows: If more than 50% of the snapshot models have correctly classified test example X then it is correctly classified. Otherwise, if $\leq 50\%$ of the votes correctly classified the example then it is considered incorrectly classified. The overall final prediction accuracy will be the percentage of correctly predicted test examples (46).

3.3. Main CNN Model Construction and Training

Finding the most efficient network parameters for a given dataset is important. It is known that parameters like batch size, number of epochs, regularization, and learning rate can significantly affect the final performance. There is no predefined method to select the ideal architecture and parameters. Therefore, we attempted multiple combinations of

settings and parameters. Our model consists of three $3 \times 3 \times 3$ convolutional layers (Conv) with a stride of $1 \times 1 \times 1$ size and ReLU activation function followed by a $2 \times 2 \times 2$ max-pooling layer with a stride of $1 \times 2 \times 2$. The first two convolutional layers have 20 filters and the third has 40. Padding was used for the convolutional layers. A fully connected layer (FC) of 64 units follows a global average pooling operation. Finally, an output with sigmoid activation is used to generate a prediction. To reduce overfitting, a dropout of 50% was added after the fully connected layer. The developed model consists of a total of 35,709 parameters.

The implementation of the model was carried out with Keras 2.3.1 using TensorFlow 1.2 as backend. Code is available on GitHub at <https://github.com/kbenahmed89/Glioma-Survival-Prediction>, accessed on 26 January 2022.

4. Experiments and Results

The experimental settings and results of this work are reported in this section. Measures include accuracy, area under the ROC curve (AUC), sensitivity, and specificity.

4.1. Parameter Exploration Results

In this section, we report some preliminary experiments performed to determine the best parameters for our dataset. Here, for simplicity, we only used a T1 contrast enhanced (T1CE) sequence as input.

4.1.1. Choice of Snapshots

The original snapshot learning method showed using the cosine annealing cyclic learning rate was most effective. It works by saving snapshots at the end of each cycle (X number of training epochs). Here, we compare the performance results of the following: The original method versus using fixed learning rate and selecting snapshots based on the training accuracy evolution. When not using cyclic learning the stochastic gradient descent optimizer was used for training the models with the starting learning rate set to 10^{-3} and momentum set to 0.9. The network state was saved at each epoch. We selected snapshots based on the training accuracy graph. The accuracy of the model starts to increase rapidly in the beginning due to the high initial learning rate. Then, it slows down and keeps increasing at a lower rate and with high stability. During the last period of training, the training accuracy has already reached 100% thus over-fitting the training data. Therefore, we decided to avoid the last set of training epochs and only focus on the earlier phase of training. To create an ensemble of snapshots, we picked the weights of every five epochs to allow a fair amount of change in the network weights. We go back 5 steps from when the training accuracy reaches 80% (epoch 30) and then proceed backwards to choose every fifth epoch as a snapshot (epochs 30, 25, 20, 15, and 10). The performance results of testing the individual five snapshots on the 46 example testing dataset are shown in Table 1 along with the results of the majority voting-based ensemble of snapshots.

To see if the original approach to snapshots might be better, we trained using a cosine annealing cyclic learning rate with a maximum learning rate of 10^{-3} . Training was done for 500 epochs and snapshots saved at the end of each of the 10 cycles, which were 50 epochs long. The majority voting-based ensemble of the last five snapshots saved at the end of the last five cycles achieved a 52.17% accuracy on the unseen test set of 46 test cases using only T1CE MRI modality. This was not as good as our approach of choosing snapshots from the training data. Of course, it is possible that a different set of parameters would result in better performance, but we had little data to test with and wanted to leave the test set alone until parameters were chosen.

Table 1. Performance results of the ensemble of five snapshots using the 3D T1CE sequence.

Method	Accuracy	Specificity	Sensitivity	AUC
Snapshot 1	63%	39%	87%	0.63
Snapshot 2	67%	74%	61%	0.67
Snapshot 3	70%	74%	65%	0.70
Snapshot 4	61%	74%	48%	0.61
Snapshot 5	59%	52%	65%	0.59
Ensemble	70%	70%	70%	0.70

Based on this experiment, we decided that our choice of snapshots based on the training graph works best for this particular dataset. Therefore, we used this snapshot choice method for the remainder of the paper.

4.1.2. Ensemble of Convolutional Neural Networks vs. Snapshot Learning

This experiment consisted of the performance comparison of an ensemble of five CNNs versus an ensemble of five snapshots from the same CNN. The CNNs have the same main architecture described in Section 3.3 but different random initializations. We decided to test at the epoch where the models reach a training accuracy of 80% to avoid overfitting. The results are presented in Table 2. The results are roughly equivalent to the ensemble of 5 fully trained models with much less training time.

Table 2. Performance results of the ensemble of five CNNs (convolutional neural networks) using the 3D T1CE sequence.

Method	Accuracy	Specificity	Sensitivity	AUC
CNN 1	63%	48%	78%	0.63
CNN 2	54%	48%	61%	0.54
CNN 3	61%	78%	43%	0.61
CNN 4	61%	65%	57%	0.61
CNN 5	57%	43%	70%	0.57
Ensemble	67%	70%	65%	0.67

4.1.3. Comparison of 3D T1CE vs. 2D T1CE MRIs

In this experiment, due to high computational costs and time required for 3D image processing, we investigated whether the third dimension added significant value over the 2D analysis of the images. We decomposed each 3D scan into 155 separate slices. Then, we manually looked at each slice and chose the one with the largest visible tumor area. Consequently, the new training set in this experiment consists of 163 images of size $240 \times 240 \times 1$. In order to increase the size of the dataset, image augmentation was also done using the same transformations explained in Section 3.1.3. We created a 2D version of the main CNN and kept the same architecture and parameters. For this dataset, after observing the training accuracy curve, we noticed that the model reaches 80% training accuracy quickly (epoch 28). Therefore, we decided to go back 5 steps from when the training accuracy reaches 90% (epoch 59) and then proceed backwards to choose every fifth epoch as a snapshot. Thus, epochs 59, 54, 49, 44, and 39 were used. The performance results of snapshots ensemble using the 2D CNN model is compared to the 3D CNN in Figure 5.

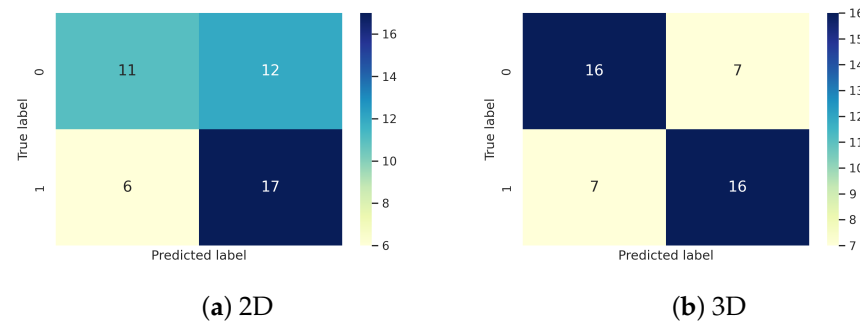


Figure 5. Confusion matrix comparison of 3D vs 2D CNN using the T1CE sequence.

4.2. Combination of Multi-Sequence Data for Survival Prediction

In this section we describe two methods attempted in order to merge the three MRI sequences (T1CE, T2, and FLAIR) to create a multi-sequence dataset for our model.

4.2.1. Ensemble of Ensembles Using T1CE, T2, and FLAIR MRIs

Before merging the three sequences, we first trained our CNN model that has the same settings and using the same training procedure described in Section 3.3, this time with the other two sequences (T2 and FLAIR) as inputs instead of T1CE. Results are reported for each sequence individually.

To create a combined decision, an ensemble of ensembles approach in which 5 snapshots from each of the three sequences T1CE, T2, and FLAIR together participate to build a final outcome prediction was used. To create each of the 5 initial ensembles, we used the 5 snapshots of each sequence that were trained individually. In a given ensemble, for each sample, one snapshot from each of the three sequences gets to vote. We repeat the process for all the 5 snapshots. We then combine the 5 voted decisions and again use majority voting at a sample level to form the final outcome.

4.2.2. Multi-Branch CNN Training with Multi-Sequence 3D MRIs

In this experiment, we combine the three sequences (T1CE, T2, and FLAIR) together to train a multi-branch 3D CNN and test it on the multi-sequence MRI test set. The architecture of the CNN is illustrated in Figure 6.

Separate CNN models operate on each sequence where each CNN has two 3D convolutional layers each followed by a max-pooling layer. Then, the results from the three models are concatenated for interpretation and ultimate prediction. The CNN was evaluated using a snapshot ensemble with max voting as explained in Section V. Here, we used a lower learning rate (10^{-4}) than in the previous experiments. Thus, the model makes small changes between epochs. Therefore, we decided to enlarge the spacing between chosen epochs. We go back 5 steps from when the training accuracy reaches 90% (epoch 183) and then proceed backwards to choose every 25th epoch as a snapshot (epochs 183, 158, 133, 108, and 83). The performance results of the multi-branch CNN are presented in Figure 7 and compared to the performance outcome of each individual sequence and to the first ensemble method described in Section 4.2.1.

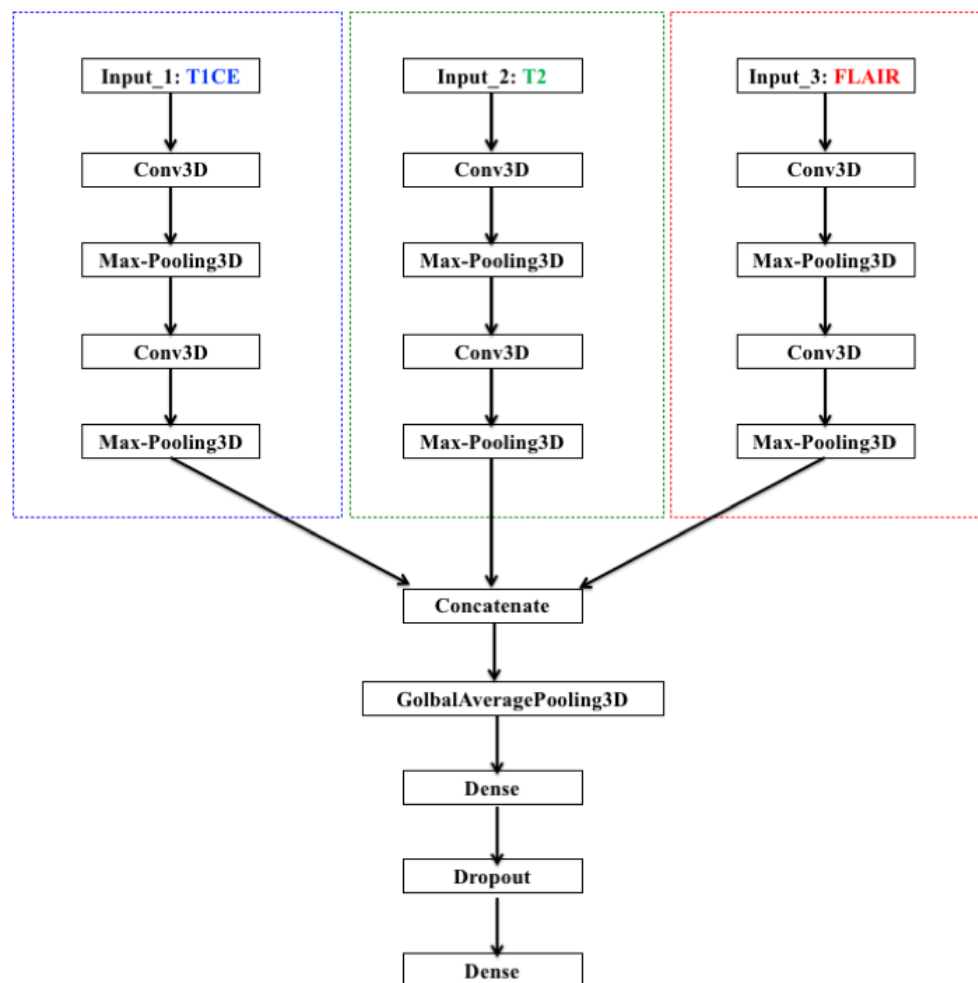


Figure 6. Multi-branch CNN architecture.

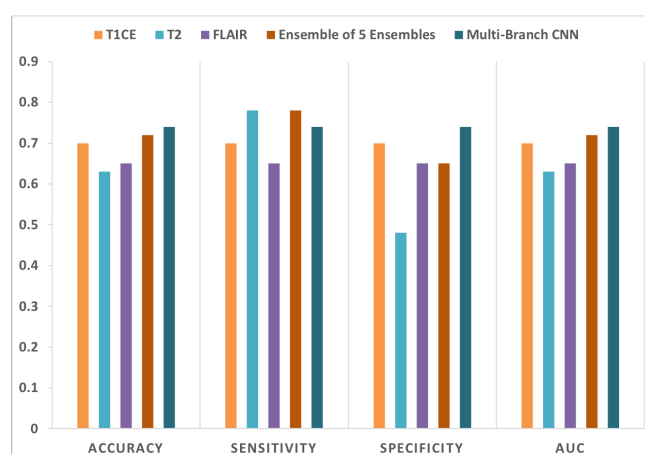


Figure 7. Performance comparison of the multi-branch CNN vs individual sequences vs ensemble of the five voted ensembles method.

4.3. Comparison with Other Methods

To further assess the prediction performance of our method, we compared our approach to the method proposed in [16]. In their paper, the authors used the BraTS 2017

training dataset with a threshold time of 18 months to classify cases into short-term and long-term survival. They extracted volumetric and location features from the 163 input images and trained a logistic regression classifier. They achieved an accuracy of 69% using 5-fold cross validation. The BraTS 2017 training dataset is a subset of the 2019 training dataset used here. To compare our results with their method, we used the same 163 cases and we also used an 18-month threshold to divide the data into two classes, short- and long-term survival. We then augmented the data and partitioned it to do the five-fold cross validation experiment. Our snapshot ensemble was used on each of the five folds. The average accuracy of the five folds was 6% higher than in [16] at 75%.

5. Discussion

As demonstrated in Table 2, the overall survival prediction performance of the ensemble of CNNs method outperformed individual CNNs. It was 67% accurate whereas the maximum we could get if we only use one individual CNN was 63%.

From Table 1, a snapshot ensemble indeed can perform as well as an ensemble of multiple CNNs and is faster and less expensive to train. Clearly, training five separate models (sequentially) is five times slower than saving snapshots of a single model. Using a snapshots ensemble, we achieved a prediction accuracy of 70%. We can also see that the performance of an ensemble is sometimes the same as only one snapshot, but it usually outperforms individual snapshots. Generally, if it can not improve the performance, it does not cause a clear decrease in accuracy.

As seen in Figure 5, comparing these results to the performance of a CNN trained using 2D images, we can conclude that even though 3D processing is expensive and time consuming, it brings relevant extra information to increase the model's performance from 61% to 70%.

Figure 7 summarizes the performance results of models trained using T1CE, T2, and FLAIR sequences standalone along with the results of the two attempted combining methods. The accuracy of the voting ensemble for T2 was 63% and for FLAIR was 65%. The 70% prediction accuracy using the T1CE sequence was slightly higher. This result probably reflects the sensitivity of T1CE to the size of the tumor core area, which is probably the most important biological predictor of overall survival in GBM patients. Experiments with combining the three sequences together were done. The first method of combination we explored was voting an ensemble of sequences. This ensemble achieved a 72% prediction accuracy. The second combining method involved creating a multi-branch CNN where each sequence was sent as input to a branch then the results were combined in the final layer. The prediction accuracy of the ensemble using this multi-branch CNN reached 74%.

The result of the two experiments with sequence combination demonstrates the potential power in the combination of multiple sequences where each contributes unique information relevant to the overall survival prediction. We can also conclude that all the tumor areas are important to determine the survival of the patient. Though perhaps the tumor core area contains the most information.

Since the BraTS testing dataset is not released by the BraTS challenge team and the validation set is only available at the time of the challenge, we were unable to directly compare our method to those applied to BraTS test/validation set. However, our prediction accuracy of 74% on an unseen testing set from BraTS 2019 suggests good generalization ability to other datasets. In addition, in the experiment conducted in Section 4.3, our method achieved a prediction accuracy of 75%, thus outperforming the approach proposed in [16], which had 69% accuracy on the same dataset.

6. Conclusions and Future Work

In summary, we demonstrated an overall survival prediction for glioblastoma patients at 74% accuracy using a multi-branch 3D convolutional neural network, where each branch was a different MR image sequence. The system classified patients into two classes (long- and short-term survival). In a comparison on the same data with the best known predictor,

our approach was 6% more accurate. An important contribution of this work is the effective use of snapshot ensembles in a novel fashion towards a solution to the limited availability of labeled images in many medical imaging datasets. Snapshots are a fast, low cost way to generate an ensemble of classifiers.

Based on recent studies [26–34], improved communication and information exchange between radiology and histopathology is needed more than ever. Many papers have shown benefits in integrating several markers for more accurate OS prediction of patients with glioma. A study [29] used deep artificial neural networks on histopathologic images of gliomas and found an association of OS with several histopathologic features, such as pseudopalisading and geographic necrosis and inflammation. Recent work [26,32,33] has shown the integrated potential of MR and histopathologic images, which has provided diagnostically relevant information for the prediction of OS in glioma. Similarly, it has been demonstrated that combining radiomics with multiomics can offer additional prognostic value [34]. Another study confirmed higher prediction accuracy when a combination of histopathologic images and genomic markers (isocitrate dehydrogenase [IDH], 1p/19q) was used [27]. Moreover, it has been shown that the addition of a radiomics model to clinical and genetic profiles improved survival prediction on glioblastoma when compared with models containing clinical and genetic profiles alone [28].

In future, we intend to create a more advanced version of our deep learning model combining radiologic-based features from MR imaging with pathologic, clinical, and genetic markers to develop more informed and better predictors of overall survival of GBM patients.

This study has the limitation of using a pre-operative dataset to perform OS prediction. This can result in leaving out several prognostic factors that are available after histological confirmation. In research, the most common prognostic factors for survival in glioblastoma patients were found to be: The age of patients, extent of resection, recursive partitioning analysis (RPA) class [35], performance status (using Karnofsky Performance Scale (KPS) or the Eastern Cooperative Oncology Group (ECOG)/World Health Organization (WHO) performance status) [36], and postoperative chemotherapy and/or radiotherapy [37].

In addition to IDH mutations [38] and 1p/19q codeletion [39], the methylation status of the O6-methyl guanine DNA methyltransferase (MGMT) gene promoter has been shown to be a strong predictor of the survival of glioblastoma patients [40]. There exists other prognostic molecular markers for which diagnostic evaluation is not routinely performed such as G-CIMP methylation, TERT promoter mutations, EGFR alterations, BRAF V600E mutations, Histone mutations, and H3K27 mutation, which can occur in histone H3.1 or H3.3 [41].

On the other hand, it may be possible to identify additional prognostic information using image analysis of the post-operative or post-treatment MRIs. A recent study has demonstrated an association between post-operative residual contrast-enhancing tumor volume in the post-surgical MRI and overall survival in newly diagnosed glioblastoma [42]. Furthermore, in our recent work [43], we showed that deep features extracted from post-treatment MRI scans, which were obtained during or after therapy, can entail relevant information for overall survival prediction. Therefore, to lend further validity to our model, we aim to combine data before and after therapy or resection, if available.

Recently, there has been increased interest in integrating non-invasive imaging techniques with clinical care in brain tumor patients. For instance, the use of amino acid PET has been shown to better identify the most biologically aggressive components of heterogeneous low and high-grade glioma [44,45]. As another future work direction, we may explore the potential of non-invasive metabolic imaging as complementary to conventional MRI to predict survival outcomes of GBM patients using convolutional neural networks.

Author Contributions: Software, K.B.A.; validation, L.O.H., D.B.G. and R.G.; data curation, K.B.A.; writing—original draft preparation, K.B.A.; writing—review and editing, L.O.H., R.G.; supervision, L.O.H., D.B.G. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: BraTS dataset can be downloaded online from the CBICA's Image Processing Portal (IPP) after requesting permission from the BraTS team. Our implementation code is available on GitHub at <https://github.com/kbenahmed89/Glioma-SurvivalPrediction>, accessed on 26 January 2022.

Conflicts of Interest: The authors declare no conflict of interest.

References

- Louis, D.N.; Perry, A.; Wesseling, P.; Brat, D.J.; Cree, I.A.; Figarella-Branger, D.; Hawkins, C.; Ng, H.; Pfister, S.M.; Reifenberger, G.; et al. The 2021 WHO classification of tumors of the central nervous system: a summary. *Neuro-Oncology* **2021**, *23*, 1231–1251.
- Ostrom, Q.T.; Gittleman, H.; Truitt, G.; Boscia, A.; Kruchko, C.; Barnholtz-Sloan, J.S. CBTRUS Statistical Report: Primary Brain and Other Central Nervous System Tumors Diagnosed in the United States in 2011–2015. *Neuro-Oncology* **2018**, *20*, iv1–iv86. doi:10.1093/neuonc/noy131.
- Stupp, R.; Taillibert, S.; Kanner, A.; Read, W.; Steinberg, D.M.; Lhermitte, B.; Toms, S.; Idhah, A.; Ahluwalia, M.S.; Fink, K.; et al. Effect of Tumor-Treating Fields Plus Maintenance Temozolomide vs Maintenance Temozolomide Alone on Survival in Patients With Glioblastoma. *JAMA* **2017**, *318*, 2306. doi:10.1001/jama.2017.18718.
- Zhou, M.; Scott, J.; Chaudhury, B.; Hall, L.; Goldgof, D.; Yeom, K.W.; Iv, M.; Ou, Y.; Kalpathy-Cramer, J.; Napel, S.; et al. Radiomics in brain tumor: image assessment, quantitative feature descriptors, and machine-learning approaches. *Am. J. Neuroradiol.* **2018**, *39*, 208–216.
- Abadi, M.; Barham, P.; Chen, J.; Chen, Z.; Davis, A.; Dean, J.; Devin, M.; Ghemawat, S.; Irving, G.; Isard, M.; et al. Tensorflow: A system for large-scale machine learning. In Proceedings of the 12th USENIX Symposium on Operating Systems Design and Implementation ({OSDI} 16), Savannah, GA, USA, 2–4 November 2016; pp. 265–283.
- Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; et al. PyTorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems*; Curran Associates Inc.: Red Hook, NY, USA, 2019; pp. 8024–8035.
- Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv* **2014**, arXiv:1409.1556.
- Szegedy, C.; Liu, W.; Jia, Y.; Sermanet, P.; Reed, S.; Anguelov, D.; Erhan, D.; Vanhoucke, V.; Rabinovich, A. Going deeper with convolutions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 1–9.
- He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
- Bernal, J.; Kushibar, K.; Asfaw, D.S.; Valverde, S.; Oliver, A.; Martí, R.; Lladó, X. Deep convolutional neural networks for brain image analysis on magnetic resonance imaging: A review. *Artif. Intell. Med.* **2019**, *95*, 64–81. doi:10.1016/j.artmed.2018.08.008.
- Walid, M. Prognostic Factors for Long-Term Survival after Glioblastoma. *Perm. J.* **2008**, *12*, 45–48. doi:10.7812/tpp/08-027.
- Bangalore Yogananda, C.G.; Shah, B.R.; Vejdani-Jahromi, M.; Nalawade, S.S.; Murugesan, G.K.; Yu, F.F.; Pinho, M.C.; Wagner, B.C.; Mickey, B.; Patel, T.R.; et al. A novel fully automated MRI-based deep-learning method for classification of IDH mutation status in brain gliomas. *Neuro-Oncology* **2020**, *22*, 402–411.
- Amian, M.; Soltaninejad, M. Multi-Resolution 3D CNN for MRI Brain Tumor Segmentation and Survival Prediction. *arXiv* **2019**, arXiv:1911.08388.
- Agravat, R.; Raval, M.S. Brain Tumor Segmentation and Survival Prediction. *arXiv* **2019**, arXiv:1909.09399.
- Rafi, A.; Ali, J.; Akram, T.; Fiaz, K.; Shahid, A.R.; Raza, B.; Madni, T.M. U-Net Based Glioblastoma Segmentation with Patient's Overall Survival Prediction. In *International Symposium on Intelligent Computing Systems*; Springer: Berlin, Germany, 2020; pp. 22–32.
- Chato, L.; Latifi, S. Machine learning and deep learning techniques to predict overall survival of brain tumor patients using MRI images. In Proceedings of the 2017 IEEE 17th International Conference on Bioinformatics and Bioengineering (BIBE), Washington, DC, USA, 23–25 October 2017; pp. 9–14.
- Hua, R.; Huo, Q.; Gao, Y.; Sui, H.; Zhang, B.; Sun, Y.; Mo, Z.; Shi, F. Segmenting Brain Tumor Using Cascaded V-Nets in Multimodal MR Images. *Front. Comput. Neurosci.* **2020**, *14*, 9.
- McKinley, R.; Rebsamen, M.; Daetwyler, K.; Meier, R.; Radojewski, P.; Wiest, R. Uncertainty-driven refinement of tumor-core segmentation using 3D-to-2D networks with label uncertainty. *arXiv* **2020**, arXiv:2012.06436.
- Wang, S.; Dai, C.; Mo, Y.; Angelini, E.; Guo, Y.; Bai, W. Automatic Brain Tumour Segmentation and Biophysics-Guided Survival Prediction. *arXiv* **2019**, arXiv:1911.08483.
- Nalepa, J.; Lorenzo, P.R.; Marcinkiewicz, M.; Bobek-Billewicz, B.; Wawrzyniak, P.; Walczak, M.; Kawulok, M.; Dudzik, W.; Kotowski, K.; Burda, I.; et al. Fully-automated deep learning-powered system for DCE-MRI analysis of brain tumors. *Artif. Intell. Med.* **2020**, *102*, 101769. doi:10.1016/j.artmed.2019.101769.

21. Menze, B.H.; Jakab, A.; Bauer, S.; Kalpathy-Cramer, J.; Farahani, K.; Kirby, J.; Burren, Y.; Porz, N.; Slotboom, J.; Wiest, R.; et al. The multimodal brain tumor image segmentation benchmark (BRATS). *IEEE Trans. Med. Imaging* **2014**, *34*, 1993–2024.
22. Wang, G.; Li, W.; Ourselin, S.; Vercauteren, T. Automatic brain tumor segmentation using cascaded anisotropic convolutional neural networks. In *International MICCAI Brainlesion Workshop*; Springer: Berlin, Germany, 2017; pp. 178–190.
23. Chollet, F.; Keras. 2015. Available online: <https://github.com/fchollet/keras> (accessed on 26 January 2022).
24. Huang, G.; Li, Y.; Pleiss, G.; Liu, Z.; Hopcroft, J.E.; Weinberger, K.Q. Snapshot ensembles: Train 1, get m for free. *arXiv* **2017**, arXiv:1704.00109.
25. Ahmed, K.B.; Hall, L.O.; Liu, R.; Gatenby, R.A.; Goldgof, D.B. Neuroimaging Based Survival Time Prediction of GBM Patients Using CNNs from Small Data. In *Proceedings of the 2019 IEEE International Conference on Systems, Man and Cybernetics (SMC)*, Bari, Italy, 6–9 October 2019; pp. 1331–1335.
26. Kickingereder, P.; Burth, S.; Wick, A.; Götz, M.; Eidels, O.; Schlemmer, H.P.; Maier-Hein, K.H.; Wick, W.; Bendszus, M.; Radbruch, A.; et al. Radiomic Profiling of Glioblastoma: Identifying an Imaging Predictor of Patient Survival with Improved Performance over Established Clinical and Radiologic Risk Models. *Radiology* **2016**, *280*, 880–889. doi:10.1148/radiol.2016160845.
27. Mobadersany, P.; Yousefi, S.; Amgad, M.; Gutman, D.A.; Barnholtz-Sloan, J.S.; Vega, J.E.V.; Brat, D.J.; Cooper, L.A. Predicting cancer outcomes from histology and genomics using convolutional networks. *Proc. Natl. Acad. Sci. USA* **2017**, *115*, 201717139. doi:10.1101/198010.
28. Bae, S.; Choi, Y.S.; Ahn, S.S.; Chang, J.H.; Kang, S.G.; Kim, E.H.; Kim, S.H.; Lee, S.K. Radiomic MRI Phenotyping of Glioblastoma: Improving Survival Prediction. *Radiology* **2018**, *289*, 797–806. doi:10.1148/radiol.2018180200.
29. Iftikhar, M.; Rathore, S.; Nasrallah, M. Analysis of microscopic images via deep neural networks can predict outcome and IDH and 1p/19q codeletion status in gliomas. *J. Neuropathol. Exp. Neurol.* **2019**, *78*, 553–553.
30. Shukla, G.; Bakas, S.; Rathore, S.; Akbari, H.; Sotiras, A.; Davatzikos, C. Radiomic features from multi-institutional glioblastoma MRI offer additive prognostic value to clinical and genomic markers: focus on TCGA-GBM collection. *Int. J. Radiat. Oncol. Biol. Phys.* **2017**, *99*, E107–E108.
31. Chaddad, A.; Daniel, P.; Desrosiers, C.; Toews, M.; Abdulkarim, B. Novel Radiomic Features Based on Joint Intensity Matrices for Predicting Glioblastoma Patient Survival Time. *IEEE J. Biomed. Health Inform.* **2019**, *23*, 795–804. doi:10.1109/jbhi.2018.2825027.
32. Rathore, S.; Chaddad, A.; Iftikhar, M.A.; Bilello, M.; Abdulkadir, A. Combining MRI and Histologic Imaging Features for Predicting Overall Survival in Patients with Glioma. *Radiol. Imaging Cancer* **2021**, *3*, e200108. doi:10.1148/rycan.2021200108.
33. Macyszyn, L.; Akbari, H.; Pisapia, J.M.; Da, X.; Attiah, M.; Pigrish, V.; Bi, Y.; Pal, S.; Davuluri, R.V.; Roccograndi, L.; et al. Imaging patterns predict patient survival and molecular subtype in glioblastoma via machine learning techniques. *Neuro-Oncology* **2015**, *18*, 417–425.
34. Chaddad, A.; Daniel, P.; Sabri, S.; Desrosiers, C.; Abdulkarim, B. Integration of radiomic and multi-omic analyses predicts survival of newly diagnosed IDH1 wild-type glioblastoma. *Cancers* **2019**, *11*, 1148.
35. Stupp, R.; Hegi, M.E.; Mason, W.P.; Van Den Bent, M.J.; Taphoorn, M.J.; Janzer, R.C.; Ludwin, S.K.; Allgeier, A.; Fisher, B.; Belanger, K.; et al. Effects of radiotherapy with concomitant and adjuvant temozolomide versus radiotherapy alone on survival in glioblastoma in a randomised phase III study: 5-year analysis of the EORTC-NCIC trial. *Lancet Oncol.* **2009**, *10*, 459–466.
36. Lamborn, K.R.; Chang, S.M.; Prados, M.D. Prognostic factors for survival of patients with glioblastoma: recursive partitioning analysis. *Neuro-Oncology* **2004**, *6*, 227–235.
37. Marina, O.; Suh, J.H.; Reddy, C.A.; Barnett, G.H.; Vogelbaum, M.A.; Peereboom, D.M.; Stevens, G.H.; Elinzano, H.; Chao, S.T. Treatment outcomes for patients with glioblastoma multiforme and a low Karnofsky Performance Scale score on presentation to a tertiary care institution. *J. Neurosurg.* **2011**, *115*, 220–229.
38. Hartmann, C.; Hentschel, B.; Wick, W.; Capper, D.; Felsberg, J.; Simon, M.; Westphal, M.; Schackert, G.; Meyermann, R.; Pietsch, T.; et al. Patients with IDH1 wild type anaplastic astrocytomas exhibit worse prognosis than IDH1-mutated glioblastomas, and IDH1 mutation status accounts for the unfavorable prognostic effect of higher age: implications for classification of gliomas. *Acta Neuropathol.* **2010**, *120*, 707–718.
39. Baumert, B.G.; Hegi, M.E.; van den Bent, M.J.; von Deimling, A.; Gorlia, T.; Hoang-Xuan, K.; Brandes, A.A.; Kantor, G.; Taphoorn, M.J.; Hassel, M.B.; et al. Temozolomide chemotherapy versus radiotherapy in high-risk low-grade glioma (EORTC 22033-26033): A randomised, open-label, phase 3 intergroup study. *Lancet Oncol.* **2016**, *17*, 1521–1532.
40. Hegi, M.E.; Diserens, A.C.; Gorlia, T.; Hamou, M.F.; De Tribolet, N.; Weller, M.; Kros, J.M.; Hainfellner, J.A.; Mason, W.; Mariani, L.; et al. MGMT gene silencing and benefit from temozolomide in glioblastoma. *N. Engl. J. Med.* **2005**, *352*, 997–1003.
41. Aquilanti, E.; Miller, J.; Santagata, S.; Cahill, D.P.; Brastianos, P.K. Updates in prognostic markers for gliomas. *Neuro-Oncology* **2018**, *20*, vii17–vii26.
42. Ellingson, B.M.; Abrey, L.E.; Nelson, S.J.; Kaufmann, T.J.; Garcia, J.; Chinot, O.; Saran, F.; Nishikawa, R.; Henriksson, R.; Mason, W.P.; et al. Validation of postoperative residual contrast-enhancing tumor volume as an independent prognostic factor for overall survival in newly diagnosed glioblastoma. *Neuro-Oncology* **2018**, *20*, 1240–1250.
43. Liu, R.; Hall, L.O.; Goldgof, D.B.; Zhou, M.; Gatenby, R.A.; Ahmed, K.B. Exploring deep features from brain tumor magnetic resonance images via transfer learning. In *Proceedings of the 2016 International Joint Conference on Neural Networks (IJCNN)*, Vancouver, BC, Canada, 24–29 July 2016; pp. 235–242.

44. Tanaka, Y.; Nariai, T.; Momose, T.; Aoyagi, M.; Maehara, T.; Tomori, T.; Yoshino, Y.; Nagaoka, T.; Ishiwata, K.; Ishii, K.; et al. Glioma surgery using a multimodal navigation system with integrated metabolic images. *J. Neurosurg.* **2009**, *110*, 163–172.
45. Arbizu, J.; Tejada, S.; Marti-Climent, J.; Diez-Valle, R.; Prieto, E.; Quincoces, G.; Vigil, C.; Idoate, M.; Zubieta, J.; Peñuelas, I.; et al. Quantitative volumetric analysis of gliomas with sequential MRI and 11 C-methionine PET assessment: patterns of integration in therapy planning. *Eur. J. Nucl. Med. Mol. Imaging* **2012**, *39*, 771–781.