

Leveraging Unlabeled Data for Glioma Molecular Subtype and Survival Prediction

Nicholas Nuechterlein, Beibin Li, Mehmet Saygin Seyfiođlu, Sachin Mehta, Patrick J Cimino, and Linda Shapiro
University of Washington, Seattle

Email: {nknuecht, beibin, msaygin, sacmehta, shapiro}@cs.washington.edu pjjc@uw.edu

Abstract—In this paper, we address two long-standing radiogenomic challenges in glioma subtype and survival prediction: (1) how to leverage large amounts of unlabeled magnetic resonance (MR) imaging data and (2) how to unite MR data and genomic data. We propose a novel application of multi-task learning (MTL) that leverages unlabeled MR data by jointly learning an auxiliary tumor segmentation task with glioma subtype prediction and that can learn from patients with and without genomic data. We analyze multi-parametric MR data from 542 patients in the combined training, validation, and testing sets of the 2018 Multimodal Brain Tumor Segmentation Challenge and somatic copy number alteration (SCNA) data from 1090 patients in The Cancer Genome Atlas' (TCGA) lower-grade glioma and glioblastoma projects. Our MTL model significantly outperforms comparable classification models trained only on labeled MR data for both IDH1/2 mutation and 1p/19q co-deletion subtype prediction tasks. We also show that embeddings produced by our MTL models improve survival predictions beyond MR or SCNA on their own. Our code is available at https://github.com/nknuecht/glioma_mtl.

I. INTRODUCTION

Gliomas make up 80% of all primary malignant brain tumors in adults [1]. The 2016 World Health Organization (WHO) criteria organizes diffuse gliomas into three broad, survival-stratifying subtypes based on the mutation status of the genes IDH1 and IDH2 and the co-deletion status of whole chromosome arms 1p and 19q [2]. Of the patients diagnosed with diffuse glioma in The Cancer Genome Atlas (TCGA), those without an IDH1/2 mutation are expected to live less than 15 months, those with an IDH1/2 mutation but without a 1p/19q co-deletion live on average just under 7 years, and patients harboring a 1p/19q co-deletion generally live nearly 11 years. Although genomic markers have become the gold standard for glioma survival stratification, such data are only attainable via costly, invasive surgery. On the other hand, magnetic resonance (MR) imaging is a cost-effective, readily available, and non-invasive method, which provides rich volumetric images of a patient's tumor. If glioma genomic subtype markers could be inferred from MR imaging or if survival prediction could be improved by complementing genomic markers with MR imaging, treatment planning, such as early enrollment in clinical trials, could be better managed.

Radiogenomics is an evolving field in medical imaging that strives to equate quantitative image features with the genomic profile of pictured tissues [3], [4]. The radiogenomic pipeline typically consists of image acquisition, image normalization, feature extraction, feature selection, and prediction using ma-

chine learning models [5]. Deep radiogenomic pipelines train end-to-end deep learning models, replacing the feature extraction, feature selection, and prediction stages with a single model [6]–[11]. In neuro-oncology, radiogenomic approaches have been used to predict WHO histopathology-defined tumor grade [12], IDH1/2 mutation status [13]–[16], 1p/19q co-deletions [11], [13], MGMT methylation [10], [11], overall and progression free survival [17]–[21] and more [22], [23].

Still, radiogenomic methods remain challenging to compare. Many studies are conducted on private datasets, and code for radiogenomic pipelines is not always released [24]. In contrast, the Multimodal Brain Tumor Segmentation Challenge (BraTS) has spearheaded open science in glioma tumor segmentation by curating large, multi-parametric glioma MR datasets [25]–[27]. The medical imaging and computer vision communities have used the BraTS datasets to train state-of-the-art segmentation models and build valuable open-source software tools [27]–[29]. Similarly, public genomic data in the TCGA have been used for groundbreaking, reproducible research, especially in glioma [30]–[32]. Of particular interest are somatic copy number alteration (SCNA) data, which describe large, contiguous deletions or duplications of DNA and have been linked to glioma molecular subtype and survival [33], [34].

Importantly, the set of patients in the 2018 BraTS dataset partially overlaps with the set of glioma patients in the TCGA [35], [36]. Labels for IDH1/2 mutations and 1p/19q co-deletions as well as overall survival (OS) are available for most patients in the intersection of these datasets, as is SCNA data. However, the majority of BraTS patients do not have accompanying data in the TCGA. Thus, training models only on BraTS MR data with genomic or survival labels leaves the bulk of these data unused and training models only on imaging data ignores rich genomic data when it is available. Multi-task learning (MTL) is a machine learning strategy that allows models to learn multiple tasks at once [37], [38]. Compared to single task learning, MTL allows models to learn richer data representations from more diverse data and labels by receiving feedback from complementary tasks [39], [40]. To allow MTL models to learn from unlabeled BraTS MR data, we exploit the recent advancements in brain tumor segmentation to assign weak segmentation labels to otherwise unlabeled MR samples.

In this paper, we propose a novel, reproducible application of MTL that (1) jointly learns tumor segmentation with IDH1/2 or 1p/19q co-deletion subtype so that samples without subtype labels can contribute to model learning and (2) unites

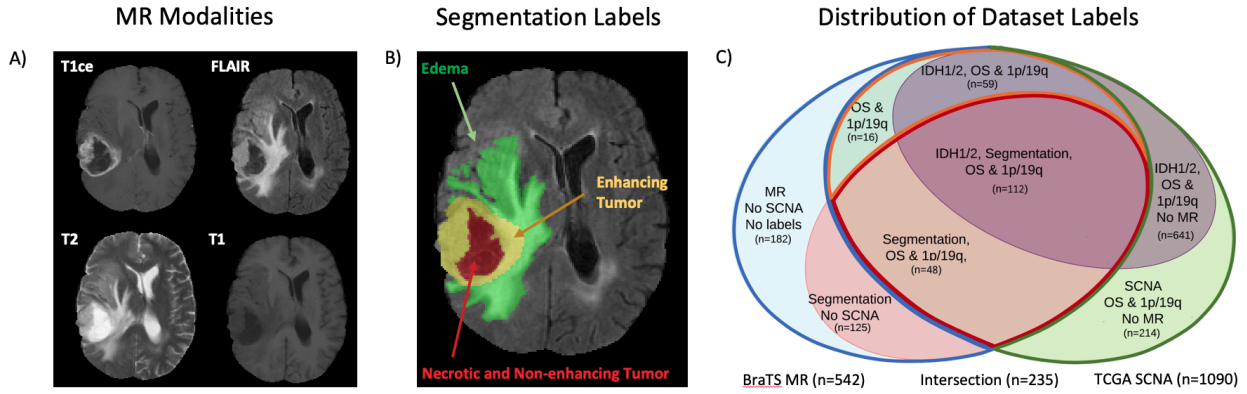


Fig. 1. **A)** Slices of four MR modalities of a WHO grade IV IDH1/2 wildtype tumor in the BraTS dataset. Each MR sample has dimension $4 \times 240 \times 240 \times 155$, where the 4 channels represent the T1ce, FLAIR, T2, and T1 MR modalities. The hallmark enhancing ring of aggressive tumors is clearly visible on the T1ce modality. **B)** A ground truth BraTS 4-class segmentation mask overlaid on the FLAIR modality. The outer class consists of peritumoral edema; the inner classes consist of necrotic tissue, non-enhancing tumor, and enhancing tumor. **C)** Distribution of dataset labels. The labeled training set is outlined with a red boundary; the validation set is outlined in a gold boundary; the unlabeled MR data leveraged by our MTL model is outlined with a blue boundary; the set of samples with SCNA data but without MR data is outlined with a green boundary. Notably, while all samples in the labeled training set and validation set have 1p/19q co-deletion and survival labels, only 69% of samples in the labeled training set have IDH1/2 labels, and only 78% of the samples in the validation set have IDH1/2 labels. Additionally, while we refer to the MR samples without subtype marker or survival labels as unlabeled, 41% of these samples do have 4-class segmentation labels.

glioma genomic and imaging data by allowing SCNA data to serve as model input alongside MR data. Our MTL model that leverages unlabeled, 4-channel MR data accurately predicts IDH1/2 mutations (AUC = 0.89) and 1p/19q co-deletions (AUC = 0.87) and outperforms comparable 3D convolutional neural networks (CNNs) trained on labeled MR imaging data alone. Training with SCNA data dramatically boosts subtype performance (AUC > 0.98) and raises the survival C-index score of survival models trained on embeddings generated by IDH1/2 prediction models from 0.719 to 0.735. To interpret our results, we train and evaluate models on individual MR modalities, visualize model gradients, and interrogate the relationship between our models’ predictions, tumor enhancement, and WHO tumor grade. To the best of our knowledge, this is the first multi-task learning strategy that unites the BraTS MR and TCGA glioma SCNA datasets.

II. METHODS

A. Dataset

Multi-parametric MR data were downloaded for 542 patients from the 2018 BraTS training, validation, and testing datasets. These MR data are 4-channel volumes composed of pre- (T1) and post-contrast (T1ce) T1-weighted modalities and T2-weighted (T2) and T2 Fluid-Attenuated Inversion Recovery (FLAIR) modalities (Figure 1A). For 285 of the 542 BraTS patients, 4-class segmentation masks that denote three different tumor compartments and a background class (Figure 1B) are given. Gene-level SCNA data were downloaded for 1090 patients in the the TCGA lower-grade (WHO grade II/III) glioma and glioblastoma (WHO grade IV) projects from the University of California Santa Cruz cancer browser¹. All 235 patients in the intersection of the BraTS MR dataset and

the TCGA glioma SCNA dataset have overall survival (OS) and 1p/19q co-deletion labels, while only 171 have IDH1/2 labels. We use the 75 patients in the BraTS 2018 validation and testing sets with SCNA data as a validation set for our models, of which 59 have IDH1/2 labels. Of the 307 samples in the BraTS dataset with neither survival nor subtype labels, 125 have ground truth segmentation labels. We assign weak segmentation labels to the other 182 samples using a public, pre-trained tumor segmentation model [41]. The distribution of labels in our dataset is shown in Figure 1C. More details are given in the supplementary materials.

B. Multi-task learning

Model: Our MTL model is illustrated in Figure 2. The backbone of our network is an open-source ESPNet-based [42] U-Net style [43] segmentation network pre-trained on the BraTS 2018 training dataset [41]. To predict molecular subtype markers, we add a branch to the bottom of the network by average pooling the output of the network’s encoder. To allow SCNA data to contribute to prediction, we pass 50-dimensional SCNA PCA embeddings through a small fully connected network and concatenate its output with the output of the average pooling step. We then pass this concatenated vector through a fully connected layer to obtain a binary IDH1/2 mutation or a binary 1p/19q co-deletion prediction. Notably, these PCA embeddings are generated from all 1,015 SCNA samples not included in the validation set (Figure 1C, outlined in red and green). Thus, we are able to leverage TCGA data that does not overlap with the BraTS dataset.

Loss Function: For subtype marker classification, we take the weighted binary cross-entropy loss \mathcal{L}_C for training samples that have subtype marker labels (Figure 1C, outlined in red). For training samples with 4-class ground truth segmentation labels (Figure 1C, red oval partially outlined in red), we

¹<https://genome-cancer.ucsc.edu/>

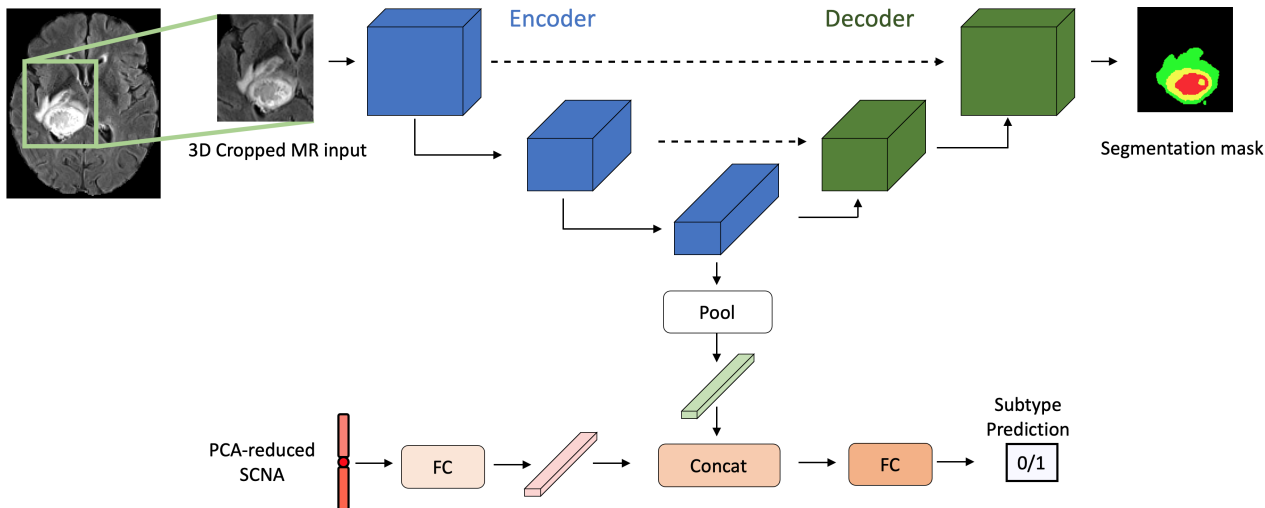


Fig. 2. Our MTL model uses the architecture of 3D-ESPNet [41] with a classification branch connected to the output of the encoder. PCA-reduced genomic SCNA data is passed through fully connected layers, concatenated with the average-pooled encoder output, and fed into a classifier to predict subtype class. The network decoder also produces a segmentation mask. We take the weighted cross-entropy loss of the subtype classification and tumor segmentation tasks. Our model accepts full brain multi-modal MR volumes as well as cropped tumor volumes (shown here). We train models on 4-channel MR data input and 1-channel, single-modality input.

take the weighted 4-class cross-entropy segmentation loss $\mathcal{L}_{S_{gt}}$ between the MTL model’s decoder’s output and the supplied segmentation mask. For unlabeled MR training samples for which 4-class segmentation labels are not available (Figure 1C, outer blue crescent), we take the weighted 2-class cross-entropy segmentation loss $\mathcal{L}_{S_{weak}}$ between the binarized output of the MTL model’s decoder and the weak binarized segmentation mask produced by the pre-trained segmentation network. To binarize these outputs, we merge all non-background segmentation classes and predict tumor vs. normal tissue. We binarize our weak segmentation masks, because the pre-trained model benchmarked higher in the BraTS competition on the whole tumor segmentation task than the individual subcompartments segmentation tasks, as is typical of models submitted to BraTS competitions [27], [41]. Finally, we define our MTL loss as

$$\mathcal{L} = \mathcal{L}_C + \lambda \mathcal{L}_{S_{gt}} + (1 - \lambda) \mathcal{L}_{S_{weak}} \quad (1)$$

where $\mathcal{L}_C = 0$ for samples without subtype marker labels, and λ is a parameter that controls the feedback from weak and ground truth segmentation labels. For samples with ground truth segmentation labels, we set $\lambda = 1$. Otherwise, we set $\lambda = 0$.

We modify $\mathcal{L}_{S_{gt}}$ when we train on 1-channel, single modality MR data. We do not consider the outer edema tumor region (Figure 1B) when we train on the T1ce modality, T1 modality, or T1ce-T1 subtraction map (described in Section III) because the edema segmentation label is difficult to identify on T1-weighted modalities. In these cases we change $\mathcal{L}_{S_{gt}}$ to the weighted 2-class cross-entropy loss between the decoder’s output and the tumor-core region comprised of the inner, non-edema classes. Conversely, the T2 and FLAIR modalities characterize the outer edema compartment better than the tumor

interior. Thus, when we train on these modalities, we change $\mathcal{L}_{S_{gt}}$ to the weighted 2-class cross-entropy loss between the decoder’s output and the binarized whole tumor mask formed by merging all non-background segmentation classes. We do not modify $\mathcal{L}_{S_{weak}}$. We discuss the segmentation labels and tumor tissue compartments in more detail in the supplementary material.

Survival Prediction: We perform survival regression analysis with the linear Cox Proportional Hazards (CPH) model, leveraging embeddings produced by the last layer in the MTL model’s classification branch [44]. To do this, we pre-train MTL models on the joint IDH1/2 mutation and tumor segmentation tasks and use the last-layer embedding to represent each training sample. We do not learn survival concurrently with subtype marker classification and tumor segmentation, because the CPH loss function requires large batch sizes that far exceed GPU memory given the size of 3D MR data (e.g., our MR raw samples are $700\times$ larger than ImageNet samples). We train our survival models on embeddings derived from the IDH1/2 mutation classification task rather than the 1p/19q co-deletion classification task, because our IDH1/2 models are more accurate and IDH1/2 mutation status stratifies survival better than 1p/19q co-deletion status.

In addition to evaluating our survival models on all patients in our validation set, we also break up our validation set by WHO 2016 molecular subtype as defined by IDH1/2 mutation and 1p/19q co-deletion status. Specifically, we consider the following three subtypes: IDH1/2 wildtype tumors, 1p/19q co-deleted tumors, and IDH1/2 mutant tumors with no 1p/19q co-deletion (1p/19q intact). We do this to evaluate the clinical utility of our models.

Input Modalities	IDH1/2 Mutation (AUC)		1p/19q Co-deletion (AUC)		Overall Survival (C-index)	
	CNN	MTL (MR)	CNN	MTL (MR)	MTL (MR)	MTL (MR+SCNA)
All (Whole Brain)	0.669	0.846	0.605	0.813	0.587	0.723
All (Cropped)	0.872	0.894	0.744	0.871	0.697	0.732
T1ce (Cropped)	0.893	0.884	0.772	0.819	0.719	0.735
FLAIR (Cropped)	0.778	0.690	0.755	0.818	0.565	0.731
T1 (Cropped)	0.731	0.738	0.727	0.757	0.645	0.728
T2 (Cropped)	0.778	0.732	0.740	0.755	0.690	0.718
T1ce-T1 (Cropped)	0.895	0.861	0.764	0.742	0.707	0.723

TABLE I
RESULTS COMPARING MTL MODELS ACROSS PREDICTION TASKS AND MR INPUT FORMAT.

III. EXPERIMENTAL RESULTS

A. Experiments

We design experiments to answer the following questions:

- 1) *Do MTL subtype marker models trained with labeled and unlabeled MR data outperform CNNs trained only on labeled MR data?* We compare our MTL model trained with unlabeled MR data and without SCNA data ($n = 467$) to 3D CNNs trained only on labeled MR data ($n = 112$ for IDH1/2 mutation prediction; $n = 160$ for 1p/19q co-deletion prediction). For a fair comparison, we match the architecture of the CNNs with that of our MTL model. Specifically, to construct our CNNs, we detach the decoder and classification branch of our MTL model and add a linear classification layer at the bottom of the network.
- 2) *Do MTL subtype marker models trained on SCNA data and MR data outperform MTL models trained only on MR data?* We train MTL models with PCA-reduced SCNA and MR data and compare them to MTL models trained only on MR data.
- 3) *Do MR and SCNA data predict survival better than either on their own?* We train linear CPH models on embeddings produced by MTL models trained on (1) MR data alone and (2) MR and PCA-reduced SCNA data and compare their results to those of a linear CPH model trained directly on PCA-reduced SCNA data.

We run these experiments on the following MR inputs:

- *All modalities (whole brain):* We concatenate all four MR volumes and use the resulting 4-channel volume as input.
- *All modalities (cropped):* We use ground truth and weak segmentation labels (described in Section II-B) to find tumor bounding boxes to crop MR volumes to the tumor boundary. We concatenate these sub-volumes and use the resulting 4-channel volume as input.
- *Single modalities (cropped):* We create 1-channel cropped tumor regions from each modality, separately.
- *T1ce-T1 subtraction map (cropped):* Subtracting the T1 volume from the T1ce volume accentuates the enhancing tumor areas thought to be more common in IDH1/2 wildtype and non-1p/19q co-deleted tumors. We use the cropped tumor region in the T1ce-T1 volume as single-channel input.

B. Evaluation Metrics and Training Details

For the IDH1/2 mutation and 1p/19q co-deletion prediction tasks, we report the average maximum AUC score for each model trained for 50 epochs over 10 trials. For survival prediction, we report the C-index, which measures the extent to which a model can properly order survival time. These are standard metrics in radiogenomic classification [7], [11], [12], [15], [16] and survival prediction [8], [18], [21].

We split our MR dataset into 467 labeled and unlabeled MR training samples (Figure 1C, large oval outlined in blue and green) and 75 MR validation samples (Figure 1C, outlined in orange). We use the remaining 855 samples with SCNA data but no MR data (Figure 1C, crescent outlined in green) to improve the SCNA PCA embeddings for the 235 samples that have SCNA and MR data, but we do not explicitly train on these 855 samples because patients rarely have SCNA data without MR data. Thus, these samples do not resemble samples our models are meant to evaluate. More training details and hyperparameter choices are described in the supplementary material.

C. IDH1/2 and 1p/19q Results

In Table I, we give IDH1/2 mutation and 1p/19q co-deletion prediction results for models trained on the set of MR inputs detailed previously. Table I shows that our MTL models almost ubiquitously outperform CNNs on the 1p/19q co-deletion task, and that the T1ce modality and T1ce-T1 subtraction map are the most effective predictors of IDH1/2 mutations. CNNs yield results mostly similar to our MTL models on the IDH1/2 classification task. Table I also shows that the most dramatic boost MTL models give is on whole brain, all modality input, where they raise classification AUC by 0.18 for IDH1/2 mutation prediction and 0.21 for 1p/19q co-deletion prediction. We point out that all of our MTL models trained with MR and SCNA data were able to classify IDH1/2 mutation and 1p/19q co-deletion status with $AUC > 0.98$. We leave these results out of Table I because they are too similar to compare.

D. Survival Results

The last columns of Table I show that, setting aside WHO tumor subtype, embeddings produced by MTL models trained on SCNA and MR data better predict glioma survival than

Input Modalities	1p/19q Co-deletion (C-index)		IDH1/2 Mutant, 1p/19q Intact (C-index)		IDH1/2 Wildtype (C-index)	
	MR	MR + SCNA	MR	MR + SCNA	MR	MR + SCNA
All (Whole Brain)	0.714	1.000	0.606	0.727	0.487	0.521
All (Cropped)	0.607	0.964	0.742	0.712	0.540	0.548
T1ce (Cropped)	0.821	0.786	0.576	0.742	0.644	0.571
FLAIR (Cropped)	0.607	0.964	0.636	0.712	0.527	0.540
T1 (Cropped)	0.643	0.893	0.606	0.636	0.535	0.544
T2 (Cropped)	0.679	0.500	0.803	0.697	0.562	0.563
T1ce-T1 (Cropped)	0.821	0.857	0.803	0.682	0.523	0.552
SCNA (PCA = 5)	0.929		0.667		0.512	

TABLE II
MTL SURVIVAL PERFORMANCE BROKEN UP OVER WHO 2016 GLIOMA SUBTYPES.

embeddings produced by MTL models trained on MR data alone. MTL models trained on T1ce data, with or without SCNA data, predict survival better than any other modality.

In Table II, we break up our survival prediction results by WHO 2016 molecular subtype as described in Section II. The most striking result is that our MTL-embedding-based survival models outperform linear CPH models trained on PCA-reduced SCNA data alone. This is strong evidence that adding MR data to SCNA data improves survival prediction.

Table II also shows that our MTL-embedding-based survival models perform exceptionally well on 1p/19q co-deleted tumors. These models perform particularly well when the embeddings they are trained on are obtained from MTL models trained on SCNA and MR input formats that contain the FLAIR modality (FLAIR, All). Our survival results also show that embeddings produced by MTL models trained on SCNA and MR data are mostly beneficial across MR input formats when predicting IDH1/2 wildtype survival, though IDH1/2 wildtype survival prediction is poor in all cases likely because all IDH1/2 wildtype tumors do extremely badly, which makes their survival difficult to order (Figure 3, red curve). The exception is that using embeddings produced by MTL models trained on only the T1ce modality gives a better survival result (C-index 0.644) over the IDH1/2 wildtype cohort than the result achieved by using embeddings produced by MTL models trained on both the T1ce modality and SCNA data (C-index 0.517). This may be indicative of the power tumor ring enhancement has. However, if this is true, we would expect the T1ce-T1 subtraction map, which also reveals tumor ring enhancement, to perform just as well, but this is not the case (C-index 0.523). In the case of IDH1/2 mutant, non-1p/19q co-deleted tumors, it is not clear whether adding SCNA to survival prediction is helpful.

IV. DISCUSSION

A. Clinical Utility

We emphasize the clinical applications of this study. For patients who are unable to undergo brain surgery, or otherwise cannot obtain their IDH1/2 mutation and 1p/19q co-deletion

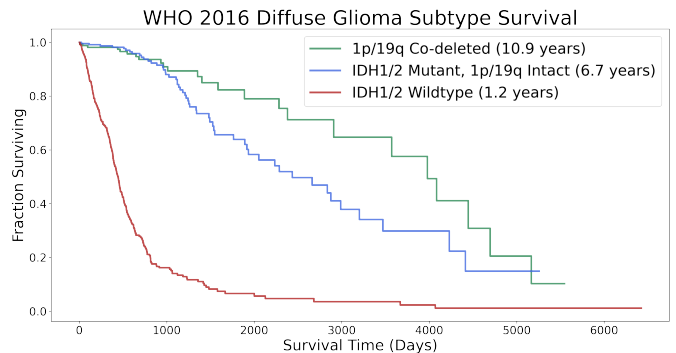


Fig. 3. Kaplan Meier survival curves for WHO 2016 molecular defined glioma subtypes from the TCGA. Median survival is given for each subtype in parentheses. IDH1/2 wildtype tumors fair the worst.

status, accurate MR-based predictions of these subtype markers place patients on dramatically different survival trajectories (Figure 3). For patients whose IDH1/2 mutation status and 1p/19q co-deletion status are known, our survival models offer the potential for sub-stratifying survival within glioma subtype. Clinically, improvements to survival stratification may lead to better treatment management, especially for predicted short-term survivors for whom early clinical trial enrollment may be recommended.

B. Tumor Ring Enhancement

We interrogate the relationship between IDH1/2 wildtype tumors and tumor ring enhancement—bright tumor tissue usually surrounding dark, necrotic tissue on the T1ce modality (Figure 4E,F). We suspect that the presence or absence of tumor ring enhancement likely drives IDH1/2 prediction in our models. To investigate this we show:

- 1) The majority of IDH1/2 wildtype tumors show tumor ring enhancement and that the majority of IDH1/2 mutant tumors show either mild or no enhancement.
- 2) Tumor IDH1/2 wildtype status correlates with tumor ring enhancement presence.
- 3) When tumor ring enhancement is present, our MTL model trained on T1ce data focuses on it.

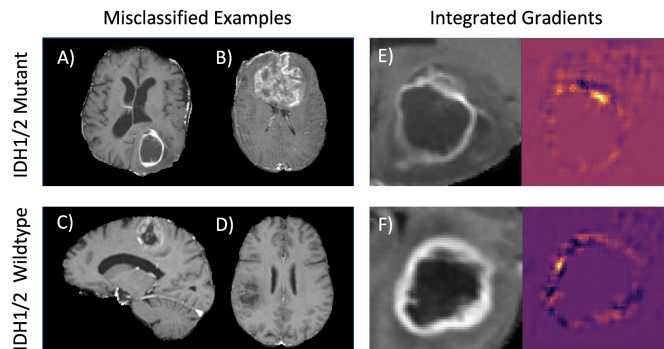
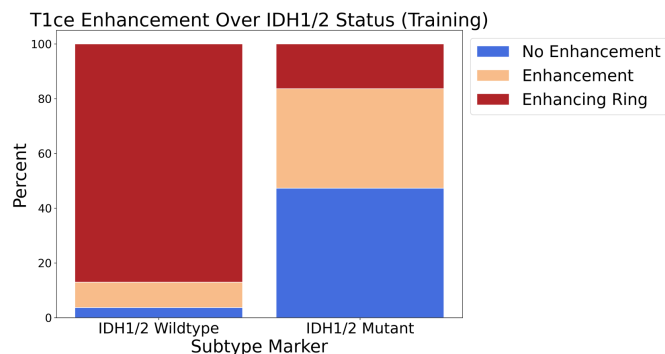


Fig. 4. **Left:** Bar chart showing the distribution of tumor enhancement over IDH1/2 status. IDH1/2 wildtype tumors predominantly display tumor ring enhancement. **Right:** Our networks appear to associate tumor ring enhancement with IDH1/2 wildtype tumors. Sample **A**) is a hyperintense IDH1/2 mutant grade IV glioma and sample **B**) is a 1p/19q co-deleted glioma. Both harbor an IDH1/2 mutation and tumor ring enhancement and are consistently misclassified as IDH1/2 wildtype gliomas. Samples **C**) and **D**) are IDH1/2 wildtype tumors that show mild and no enhancement, respectively, and are misclassified as IDH1/2 mutants. Sample **E**) is an IDH1/2 mutant misclassified as an IDH1/2 wildtype. Sample **F**) is a correctly classified IDH1/2 wildtype. Integrated gradients in the images to the right of **E**) and **F**) show that our T1ce-based MTL models put emphasis on tumor ring enhancement.

- 4) Our model’s IDH1/2 validation predictions score well when considered as tumor ring enhancement predictions.

First, we plot the distribution of tumor ring enhancement across IDH1/2 mutant and wildtype tumors in the labeled IDH1/2 training set (Figure 4, left panel). We see that tumors with tumor ring enhancement are indeed far more likely to be IDH1/2 wildtype tumors whereas tumors with mild or no enhancement are more likely to have IDH1/2 mutations. Second, we find that the Pearson correlation coefficient ($R = 0.791, p < 10e^{-14}$) between IDH1/2 wildtype status and enhancing ring presence indicates that they correlate.

Third, to see whether our IDH1/2 MTL model trained on the T1ce modality focuses on tumor ring enhancement when it is present, we use integrated gradients, a gradient attribution method, to visualize the regions of the tumor on which our model concentrates [45]. Integrated gradients pictured in Figures 4E and 4F show that this is the case. We see more evidence that this model associates tumor ring enhancement with IDH1/2 wildtype tumors when we examine the mistakes this model makes. Examples where this model misclassifies IDH1/2 mutant tumors with ring enhancement as IDH1/2 wildtype tumors are given in Figures 4A and 4B; on Figures 4C and 4D we see that our model mistakes IDH1/2 wildtype tumors without tumor ring enhancement for IDH1/2 mutant tumors. Finally, we calculate the AUC score (0.955) between our model’s IDH1/2 predictions and tumor ring enhancement presence labels. That the AUC score is higher when measured against ring enhancement labels than IDH1/2 mutation labels strongly suggests that this model associates IDH1/2 wildtype tumors with tumor ring enhancement to the point of overfitting.

We investigate the relationship between 1p/19q co-deleted tumors and tumor ring enhancement in the supplementary material, although this effort is hampered by the low count of 1p/19q co-deleted tumors in the labeled training set (14/160) and validation set (14/75).

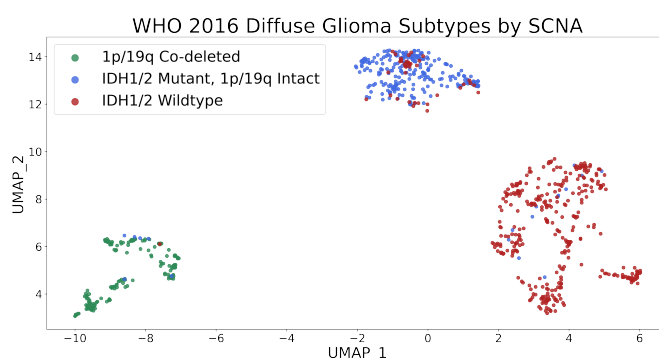


Fig. 5. A UMAP projection of 1090 SCNA samples produces mostly WHO molecular subtype-specific clusters.

C. Additional Performance Comments

We discuss two important results. First, we note that adding SCNA data to our MTL models pushes the AUC score over 0.98 for both subtype marker prediction tasks. A UMAP projection [46] of the entire SCNA TCGA dataset (Figure 5) shows that WHO molecular subtypes mostly cluster together and therefore are likely predictive of subtype markers on their own. Thus, MR imaging may be unnecessary for subtype marker prediction in the rare setting where patients have SCNA data but no direct knowledge of their IDH1/2 mutation or 1p/19q co-deletion status. Nonetheless, the observation that SCNA is correlated with WHO glioma subtype makes it more impressive that our MTL-based survival models outperform survival models trained on raw PCA-reduced SCNA data alone given how well glioma subtype stratifies survival (Figure 3).

Second, we observe that adding unlabeled data to the 4-channelled whole brain MR input boosts subtype marker prediction far more than other MR input formats (Table I, All (Whole Brain)). We suspect that this is because we initialize the encoder and decoder components of our MTL network with the weights from a tumor segmentation network pre-trained on the BraTS data. Because the pre-trained network

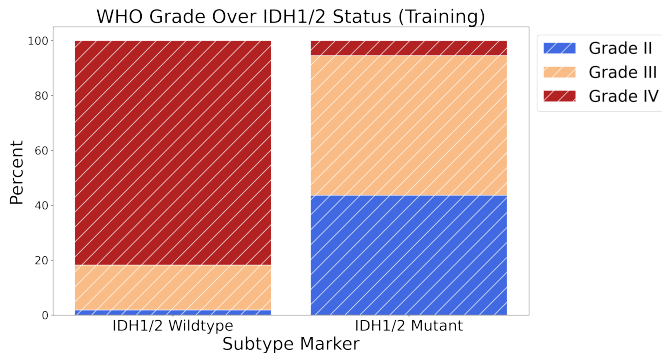


Fig. 6. Bar chart showing the distribution of WHO grade over IDH1/2 status in our labeled training set. IDH1/2 wildtype tumors are known to be predominantly grade IV tumors [2].

was trained on 4-modality data at the same resolution we use for our whole brain training input, whole brain MTL models give reasonable segmentation outputs and thus are likely stabilized by the auxiliary segmentation task. On the other hand, the higher resolution of whole brain input coupled with the fact that CNNs are limited to labeled MR training data likely makes whole brain input particularly difficult for single task learning.

D. Comparison to Previous Work

We do not give direct comparisons to previously reported results, because our goal was not outperform others. Instead, we endeavored to test the advantage of using unlabeled MR over supervised CNNs and did not want optimizations, such as network architecture, to obscure the benefit of our strategy. We also do not formally report our segmentation results, because segmentation is performed only to increase classification and survival prediction performance, not to compete with dedicated segmentation models. Our MTL model trained on the 1p/19q co-deletion task with 4-channel cropped MR input gives the best segmentation results (Dice score 0.70). Although this is substantially worse than our pre-trained network’s reported performance (Dice score 0.85) and that of the best BraTS models (Dice scores ≈ 0.90), using segmentation loss in our MTL models accomplishes our goal of boosting the subtype marker and survival tasks’ performances [27], [28].

E. Future Directions

Natural extensions of this work include further model optimization, adding concurrent survival learning to our MTL models, and conducting a deeper investigation into our work’s clinical implications. Upgrading the pre-trained segmentation model we use to a model that benchmarked higher on a BraTS challenge constitutes a simple improvement to this study. We chose an ESPNet-based encoder-decoder structure for our MTL model backbone, because ESPNet is light-weight, fast to finetune, and delivers performance competitive with other off-the-shelf segmentation models [42].

Ideally, we would like to learn overall survival jointly with glioma subtype. However, our survival models must be trained

separately, because the CPH loss function requires $O(n^2)$ comparisons for each pair of data points. For large 3D MR data, this requires too much memory. Siamese networks, triplet loss, and other network designs allow training pairwise losses in mini-batch fashion, but integrating CPH loss with mini-batches degrades the prediction performance when the number of samples is much larger than the batch size [47]. Future studies are needed to design better deep learning models for survival prediction in this complex setting.

We showed that IDH1/2 wildtype status is tightly correlated with the presence of tumor ring enhancement, and thus that accurate prediction of IDH1/2 mutations results in accurate prediction of the presence of tumor ring enhancement. Therefore, it may be more prudent to predict IDH1/2 status for tumors known to have tumor ring enhancement, because this will control for such correlation. Similarly, our analysis could be improved by controlling for WHO tumor grade. Like tumor ring enhancement, WHO grade (grade II/III vs. grade IV) is correlated with IDH1/2 wildtype status ($R = 0.742, p < 10e^{-11}$) and has a AUC score of 0.95 when evaluated with our MTL model’s IDH1/2 predictions. Figure 6 shows that IDH1/2 mutant tumors in our training set are mostly WHO grade II/III, while IDH1/2 wildtype tumors are mostly WHO grade IV. Because tumor grade is known to stratify survival, not only should we evaluate our survival models separately on each subtype, we should also break each subtype down by grade. However, sample size and class imbalance are obstacles to both of these conditions. Tumor ring enhancing IDH1/2 mutant tumors are rare, for example, as are WHO grade IV IDH1/2 mutant tumors. Larger cohorts are necessary for these analyses.

V. CONCLUSION

The ability of MTL to learn from unlabeled imaging data is a promising avenue for expanding existing datasets. Further, models capable of learning from genomic and imaging data create the potential for quickly piecing together diverse datasets that historically have been analyzed independently. This is especially important for public datasets where new methods can be collectively developed and reproduced. To this end, we merged the 2018 BraTS MR dataset and the TCGA lower-grade (WHO grade II/III) glioma and glioblastoma (WHO grade IV) SCNA datasets. Our primary contribution is the novel application of MTL that jointly learns classification tasks and tumor segmentation, allowing unlabeled MR data to contribute to model learning. We show that using MTL to leverage unlabeled MR data improves subtype marker prediction and, by training on MTL embeddings, survival prediction as well. When SCNA data is used in conjunction with MR data, our results further improve. Our models have the potential to aid patient management, especially for predicted short-term survivors.

ACKNOWLEDGMENTS

Nicholas Nuechterlein is supported by the National Science Foundation under Grant DGE-1762114.

REFERENCES

- [1] M. L. Goodenberger and R. B. Jenkins, "Genetics of adult glioma," *Cancer genetics*, vol. 205, no. 12, 2012.
- [2] D. N. Louis *et al.*, "The 2016 world health organization classification of tumors of the central nervous system: a summary," *Acta neuropathologica*, vol. 131, no. 6, 2016.
- [3] R. J. Gillies *et al.*, "Radiomics: images are more than pictures, they are data," *Radiology*, vol. 278, no. 2, 2016.
- [4] A. M. Rutman *et al.*, "Radiogenomics: creating a link between molecular diagnostics and diagnostic imaging," *European journal of radiology*, vol. 70, no. 2, 2009.
- [5] V. Kumar *et al.*, "Radiomics: the process and the challenges," *Magnetic resonance imaging*, vol. 30, no. 9, 2012.
- [6] R. Fukuma *et al.*, "Prediction of idh and tert promoter mutations in low-grade glioma from magnetic resonance images using a convolutional neural network," *Scientific Reports*, vol. 9, no. 1, 2019.
- [7] Z. Li *et al.*, "Deep learning based radiomics (dlr) and its usage in noninvasive idh1 prediction for low grade glioma," *Scientific reports*, vol. 7, no. 1, 2017.
- [8] J. Lao *et al.*, "A deep learning-based radiomics model for prediction of survival in glioblastoma multiforme," *Scientific reports*, vol. 7, no. 1, 2017.
- [9] K. Chang *et al.*, "Residual convolutional neural network for the determination of idh status in low-and high-grade gliomas from mr imaging," *Clinical Cancer Research*, vol. 24, no. 5, 2018.
- [10] P. Korfiatis *et al.*, "Residual deep convolutional neural network predicts mgmt methylation status," *Journal of digital imaging*, vol. 30, no. 5, 2017.
- [11] P. Chang *et al.*, "Deep-learning convolutional neural networks accurately classify genetic mutations in gliomas," *American Journal of Neuroradiology*, vol. 39, no. 7, 2018.
- [12] C. Su *et al.*, "Radiomics based on multicontrast mri can precisely differentiate among glioma subtypes and predict tumour-proliferative behaviour," *European radiology*, vol. 29, no. 4, 2019.
- [13] Y. Matsui *et al.*, "Prediction of lower-grade glioma molecular subtypes using deep learning," *Journal of Neuro-Oncology*, vol. 146, no. 2, 2020.
- [14] H. Arita *et al.*, "Lesion location implemented magnetic resonance imaging radiomics for predicting idh and tert promoter mutations in grade ii/iii gliomas," *Scientific reports*, vol. 8, no. 1, 2018.
- [15] X. Zhang *et al.*, "Radiomics strategy for molecular subtype stratification of lower-grade glioma: detecting idh and tp53 mutations based on multimodal mri," *Journal of Magnetic Resonance Imaging*, vol. 48, no. 4, 2018.
- [16] Z.-C. Li *et al.*, "Multiregional radiomics profiling from multiparametric mri: Identifying an imaging predictor of idh1 mutation status in glioblastoma," *Cancer medicine*, vol. 7, no. 12, 2018.
- [17] K. E. Emblem *et al.*, "Machine learning in preoperative glioma mri: Survival associations by perfusion-based support vector machine outperforms traditional mri," *Journal of magnetic resonance imaging*, vol. 40, no. 1, 2014.
- [18] P. Kickingereder *et al.*, "Radiomic profiling of glioblastoma: identifying an imaging predictor of patient survival with improved performance over established clinical and radiologic risk models," *Radiology*, vol. 280, no. 3, 2016.
- [19] —, "Radiomic subtyping improves disease stratification beyond key molecular, clinical, and standard imaging characteristics in patients with glioblastoma," *Neuro-oncology*, vol. 20, no. 6, 2018.
- [20] X. Liu *et al.*, "A radiomic signature as a non-invasive predictor of progression-free survival in patients with lower-grade gliomas," *NeuroImage: Clinical*, vol. 20, 2018.
- [21] Q. Li *et al.*, "A fully-automatic multiparametric radiomics model: towards reproducible and prognostic imaging signature for prediction of overall survival in glioblastoma multiforme," *Scientific reports*, vol. 7, no. 1, 2017.
- [22] P. Kickingereder *et al.*, "Large-scale radiomic profiling of recurrent glioblastoma identifies an imaging predictor for stratifying anti-angiogenic treatment response," *Clinical Cancer Research*, vol. 22, no. 23, 2016.
- [23] P. Kickingereder, D. Bonekamp *et al.*, "Radiogenomics of glioblastoma: machine learning-based classification of molecular characteristics by using multiparametric and multiregional mr imaging features," *Radiology*, vol. 281, no. 3, pp. 907–918, 2016.
- [24] J. E. Park *et al.*, "A systematic review reporting quality of radiomics research in neuro-oncology: toward clinical utility and quality improvement using high-dimensional imaging features," *BMC cancer*, vol. 20, no. 1, 2020.
- [25] B. H. Menze *et al.*, "The multimodal brain tumor image segmentation benchmark (brats)," *IEEE transactions on medical imaging*, vol. 34, no. 10, 2014.
- [26] S. Bakas, H. Akbari *et al.*, "Advancing the cancer genome atlas glioma mri collections with expert segmentation labels and radiomic features," *Scientific data*, vol. 4, 2017.
- [27] S. Bakas *et al.*, "Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the brats challenge," *arXiv preprint arXiv:1811.02629*, 2018.
- [28] K. Kamnitsas *et al.*, "Deepmedic for brain tumor segmentation," in *International workshop on Brainlesion: Glioma, multiple sclerosis, stroke and traumatic brain injuries*. Springer, 2016.
- [29] S. Rathore *et al.*, "Brain cancer imaging phenomics toolkit (brain-captk): an interactive platform for quantitative analysis of glioblastoma," in *International MICCAI Brainlesion Workshop*. Springer, 2017.
- [30] C. G. A. R. Network *et al.*, "Comprehensive genomic characterization defines human glioblastoma genes and core pathways," *Nature*, vol. 455, no. 7216, p. 1061, 2008.
- [31] R. G. Verhaak *et al.*, "Integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterized by abnormalities in pdgfra, idh1, egfr, and nf1," *Cancer cell*, vol. 17, no. 1, pp. 98–110, 2010.
- [32] H. Noshmeh *et al.*, "Identification of a cpg island methylator phenotype that defines a distinct subgroup of glioma," *Cancer cell*, vol. 17, no. 5, pp. 510–522, 2010.
- [33] P. J. Cimino *et al.*, "Multidimensional scaling of diffuse gliomas: application to the 2016 world health organization classification system with prognostically relevant molecular subtype discovery," *Acta neuropathologica communications*, vol. 5, no. 1, pp. 1–14, 2017.
- [34] P. J. Cimino, L. McFerrin *et al.*, "Copy number profiling across glioblastoma populations has implications for clinical trial design," vol. 20, no. 10, pp. 1368–1373, 2018.
- [35] S. Bakas *et al.*, "Segmentation labels and radiomic features for the pre-operative scans of the tcga-1gg collection," *The Cancer Imaging Archive*, vol. 286, 2017.
- [36] S. Bakas, H. Akbari *et al.*, "Segmentation labels and radiomic features for the pre-operative scans of the tcga-gbm collection. the cancer imaging archive," *Nat Sci Data*, vol. 4, p. 170117, 2017.
- [37] R. Caruana, "Multitask learning," *Machine learning*, vol. 28, no. 1, pp. 41–75, 1997.
- [38] T. Evgeniou *et al.*, "Regularized multi-task learning," in *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, 2004.
- [39] K. He *et al.*, "Mask r-cnn," in *Proceedings of the IEEE international conference on computer vision*, 2017.
- [40] E. Hajiramezani *et al.*, "Bayesian multi-domain learning for cancer subtype discovery from next-generation sequencing count data," in *Advances in Neural Information Processing Systems*, 2018, pp. 9115–9124.
- [41] N. Nuechterlein *et al.*, "3D-ESPNet with pyramidal refinement for volumetric brain tumor image segmentation," in *International MICCAI Brainlesion Workshop*. Springer, 2018.
- [42] S. Mehta *et al.*, "ESPNet: Efficient spatial pyramid of dilated convolutions for semantic segmentation," in *Proceedings of the european conference on computer vision (ECCV)*, 2018.
- [43] O. Ronneberger *et al.*, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015.
- [44] D. R. Cox, "Regression models and life-tables," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 34, no. 2, pp. 187–202, 1972.
- [45] M. Sundararajan *et al.*, "Axiomatic attribution for deep networks," in *Proceedings of the 34th International Conference on Machine Learning-Volume 70*. JMLR. org, 2017.
- [46] L. McInnes *et al.*, "Umap: Uniform manifold approximation and projection for dimension reduction," *arXiv preprint arXiv:1802.03426*, 2018.
- [47] G. Koch *et al.*, "Siamese neural networks for one-shot image recognition," in *ICML deep learning workshop*, vol. 2. Lille, 2015.