CS230

Spatially-Resolved Deep Radiomics Classification of Tumor-Immune Landscapes

Healthcare Application Project: Final Report

Dileep D. Monie*
Department of Computer Science
Stanford University
dmonie@stanford.edu

Abstract

Immunotherapies such as immune checkpoint inhibitors (ICIs) and chimeric antigen receptor (CAR) T cells have made significant strides in treating a variety of cancers. Developing these treatments for neuro-oncology has been difficult, however, because of the difficulty in assessing tumor response noninvasively. Deep radiomics models that use radiographic image features to predict cellular states are a promising solution. Since brain images are not homogeneous for any given cell type, the challenge is spatial resolution of these predictions. In this study, we built a spatially-resolved deep radiomics binary classifier of T cell tumor infiltration by training a modified ResNet50 architecture with magnetic resonance (MR) brain image features and stereotactic biopsy bulk RNA sequencing (RNA-seq) labels. Our model can predict whether or not CD3δ transcripts per kilobase million (TPMs) are above or below the population median with a binary accuracy of 0.57. Further engineering of our input features and model architecture, along with hyperparameter tuning, is needed to deploy this as a clinical decision aid for immunotherapy.

1 Introduction

Glioblastoma is the most aggressive and lethal brain tumor in adults. Median survival after diagnosis is about 15 months with the current standard of care, which entails surgical resection with adjuvant chemotherapy and radiation. Emerging immunotherapies such as immune checkpoint inhibitors (ICIs) and chimeric antigen receptor (CAR) T cells offer patients hope. These immunotherapies work by increasing inflammation at the tumor site, which appears to be disease progression on magnetic resonance (MR) imaging. Experienced radiologists cannot distinguish between this pseudoprogression and true progression and, therefore, an invasive biopsy is required to make a treatment decision. This project aims to address this problem by building a spatially-resolved binary classifier that can make this distinction using just MR images.

The presence of tumor-infiltrating lymphocytes (TILs) are of particular interest to assessing immunotherapy responses. Successful treatment with checkpoint blockade or CAR-T cells are expected to increase the number of TILs, particularly T cells. CD3 is a pan-T cell marker made up of multiple subunits: a CD3γ, a CD3δ, and two CD3ϵ chains [1]. We can measure the expression of these genes using bulk RNA sequencing (RNA-seq) on tumor biopsies. For our dataset, CD3δ had the highest median and mean transcripts per kilobase million (TPMs). Therefore we designed the classifier to

*https://profiles.stanford.edu/dmonie

CS230: Deep Learning, Spring 2021, Stanford University, CA.

predict whether or not an image feature would have CD3δ expression above or below the population median.

Given that brain images are not homogeneous for T cells or any other cell type, the challenge here is spatial resolution of these predictions. Our classifier looks at the region just around the biopsy site, rather than the full image. We trained a ResNet50 model that can predict whether or not CD3δ expression is above or below the population median TPMs in a 17x17 pixel MR image region with a binary accuracy of 0.57.

2 Related work

High dimensional radiomics models have been developed to predict the molecular heterogeneity between patients, but few have attempted to capture the robust spatial heterogeneity in these molecular markers within a single tumor. These attempts used earlier iterations of our dataset and are published in [2], [3], [4], and [5]. Data preprocessing approaches are discussed in [6], [7], [8], and [9]. We published a proof of concept of our "shallow" machine learning approach in [10], shown for tumor cell proliferation and death markers in **Figure 1**. Qualitatively, once further developed, we will deploy our deep learning model to generate similar multi-genic pheno-region maps of immune cell molecular signatures superimposed on brain MR images.

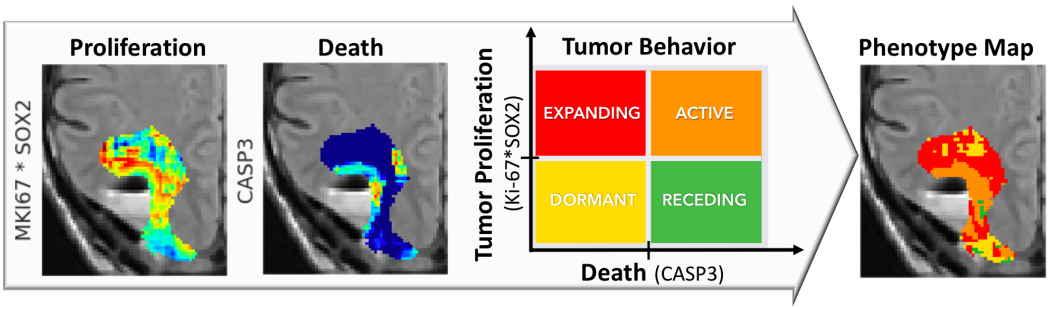

Figure 1: Defining phenotypically distinct regions (pheno-regions) provide a means to visualize the relative proliferative/death activity of different regions of the tumor.

A number of trained immune radiomics models for a variety cancers, all lacking spatial resolution, are reviewed in [10]. Additional datasets, also without spatial resolution, are publicly available through The Cancer Imaging Archive (TCIA; `https://www.cancerimagingarchive.net/`).

Our approach here is unique because it combines deep leaning with spatially-resolved radiomics. The deep learning pipeline consists of three key steps including data selection, preprocessing, and radiomics model training with validation.

3 Dataset and Features

We are using unpublished, de-identified glioblastoma patient data from the Mayo Clinic Mathematical Neuro-Oncology Laboratory, collected in collaboration with the Barrow Neurological Institute, Columbia University, and City of Hope. These patients are enrolled in a clinical trial examining the safety on efficacy of CAR-T cell therapy directed against IL-13Rα2, with or without concurrent ICI. This dataset contains 150 spatially-resolved biopsies from 45 patients, each with up to 9 pretreatment coregistered MR image sequences: diffusion tensor imaging (DTI) with fractional anisotropy (FA), DTI with mean diffusivity (MD), echo planar (EPI-C), fluid-attenuated inversion recovery (FLAIR), relative cerebral blood volume (rCBV), standardized rCBV, T1-weighted, T1-weighted with gadolinium contrast enhancement (T1Gd), or T2-weighted. An example T1Gd image slice is shown in **Figure 2A** with biopsy site marked and the corresponding 17x17 input feature channel with biopsy coordinates as the center pixel is shown in **Figure 2B**.

All 9 of the 17x17 were stacked into channels to generate a 17x17x9 input feature for each biopsy. An alternative 3-channel faux RGB image composed of z-slices adjacent to the biopsy coordinates

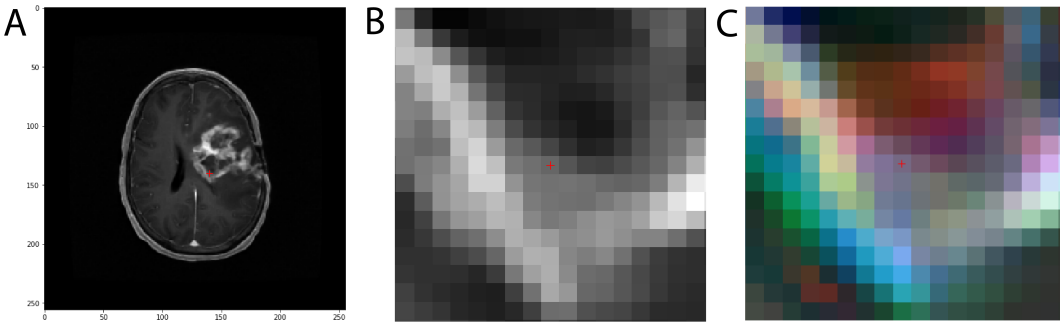

Figure 2: (A) An example slice of a 256x256 T1Gd image. (B) The corresponding T1Gd channel of the 17x17 input feature. (C) A 3-channel faux RGB image composed of z-slices adjacent to the biopsy coordinates. The same biopsy coordinates are marked by a red + in each image.

that could be stacked into channels to generate a 17x17x27 input feature is shown in **Figure 2C** for illustration, but was not used in the current model due to technical challenges.

Each biopsy has associated RNA-seq data used to detect immune cell signatures. Median expression of the pan-T cell marker CD3δ is 1.03 TPMs in our full dataset ($n = 150$). The Boolean True served as the model output label for examples with expression greater than this median ($n = 75$). All other features were labeled False ($n = 75$). Given that we have 150 examples, quite small by deep learning standards, we decided to randomly shuffle and split it up into a 60% training set ($n = 90$), a 20% dev set ($n = 30$), and a 20% test set ($n = 30$).

While beyond the scope of the current study, there are about 50 repeat MR images acquired on different scanners within a few days of each other that could be used to address portability issues and increase robustness of the trained model.

4 Methods

For this project, we used a Keras implementation of a ResNet50 model architecture as described in [15]. This is a convolutional neural network (CNN) with a depth of 50 layers made up of residual blocks. These residual blocks contain skip connections that provide a shortcut to downstream activation functions, in this case the rectified linear unit (ReLU):

$$f(x) = \begin{cases} 0 & x \leq 0 \\ x & x > 0 \end{cases}$$

The default input shape designed for larger 3-channel color images was modified to accommodate our (17, 17, 9) input features. For the output, we used 2 fully connected layers leading up to a single node with sigmoid activation because we wanted a binary classifier. ReLU was used as the activation function for all other layers. This deep neural network has 24,164,097 total parameters, 24,110,977 of which were trainable and 53,120 were non-trainable. The full network architecture is summarized in the accompanying Jupyter notebook.

While we monitored a number of metrics (described below), none of these were differentiable and thus not suitable to be a loss function. We therefore used binary cross-entropy or log loss:

$$H_p(q) = -\frac{1}{N}\sum_{i=1}^{N} y_i * log(p(y_i)) + (1 - y_i) * log(1 - p(y_i))$$

where y_i is our label (1 for CD3δ TPM $>$ median and 0 for CD3δ TPM $\leq$ median) and $p(y_i)$ is the probability that CD3δ TPM $>$ median for a given biopsy feature i.

5 Experiments/Results/Discussion

Given the novelty of our dataset and our application, most of our efforts were focused on feature engineering. We tried many different shapes of the input features to balance a small biopsy radius

with providing enough information to the model. Increasing or decreasing our radius did not provide any benefit. Thus we settled on an 8 pixel radius because that is approximately 2-3 times the error in our stereotactic biopsy registration. Data augmentation by including multiple z-slices (**Figure 2C**) and rotating or flipping images was not technically feasible in the time allotted. Similarly, using floating point TPM labels and multi-genic labels were attempted but unsuccessful due to time and technical limitations.

We began our model experimentation by using AutoKeras to tune both the model and hyperparameters. This failed to yield a useful classification model under given time constraints so a manual, rational approach was employed. For our ResNet50 model, we used an Adam optimizer because current best practices suggest that it may be best for our application [13]. The Adam variant AMSgrad and stochastic gradient decent (SGD) were also tried but did not improve our model. The learning rate was the first hyperparameter that we adjusted, starting from the default 0.001. Increasing to 0.01 or 0.1 caused the dev loss to increase, causing high variance. Decreasing to 0.0001 or 0.00001 did not improve the model so we reverted to the default learning rate. We played around with the batch size, ranging from 1 to 90 (100% of training set). This had little impact on our outcomes but a batch size of 9 (10% of training set) yielded the most reproducible results. In some experiments, the variance increased in later epochs that could be avoided with early stopping. This is not ideal because it lacks orthagonality so cost function optimization and regularization may be better.

We also started building ResNet152 and DenseNet201 models but abandoned these due to time limitations [14]. These are highly promising as a future direction.

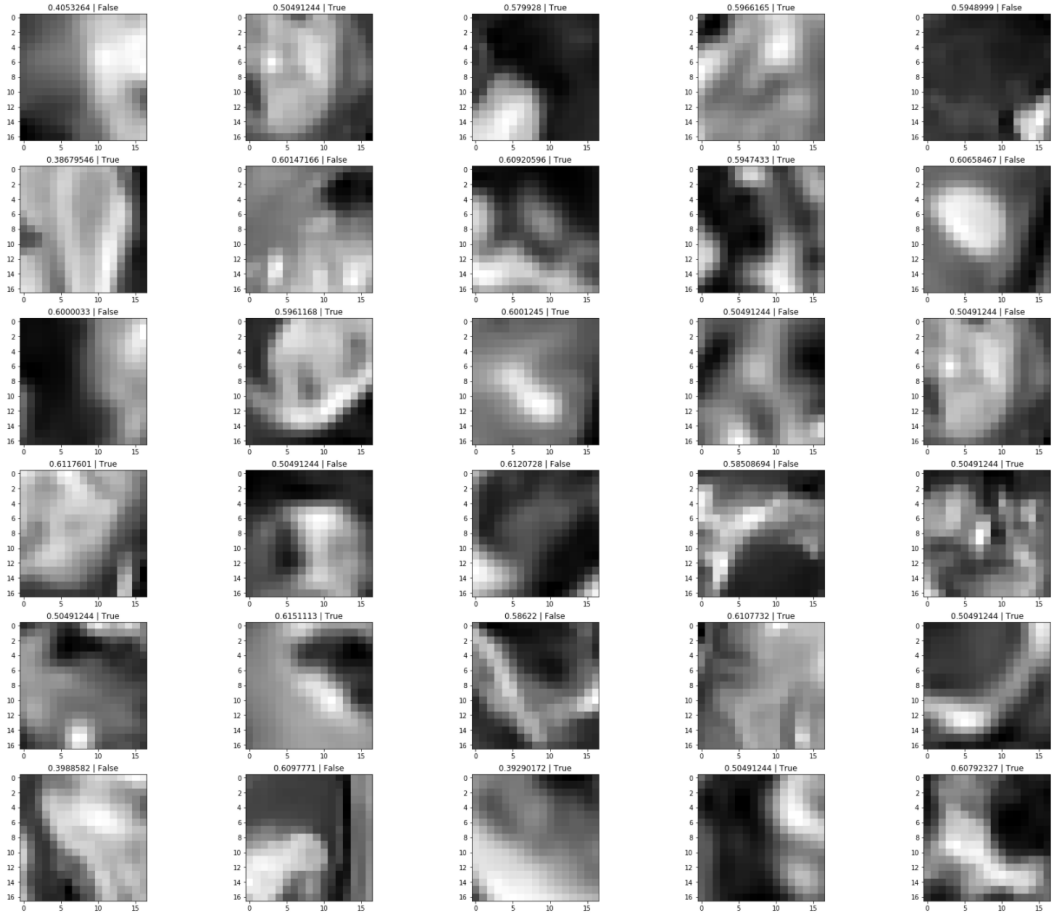

Figure 3: Predictions (represented as the probability that CD3δ TPM $>$ median for the given biopsy feature) along with labels and T1Gd channel image of the 17x17 features for the test set ($n = 30$).

For metrics, binary accuracy is useful because we have a balanced dataset with equal numbers of both classes. Our final model had a binary accuracy of 0.5667 on the test set. An F1 score (harmonic mean

of precision and recall) may be better in general, but has been removed from the latest Keras and a workaround was not feasible in the given time. We also used the area under the receiver operating characteristic curve (AUC) as a secondary metric. AUC maximizes the true positive rate (TPR) while simultaneously minimizing the false positive rate (FPR). The final model had an AUC of 0.5136 on the test set (curve not shown due to a system failure). The predictions along with labels and T1Gd channel of the 17x17 features for the entire test set are shown in **Figure 3**.

This task is beyond human capabilities with experienced radiologists having a 50% accuracy. So while these results are not great, they suggest that the problem may be feasible with further data and model engineering combined with more extensive hyperparameter tuning.

6 Conclusion/Future Work

In summary, we have built the first spatially-resolved deep radiomics classifier of brain tumor T cell infiltration. We engineered biopsy image features comprised of 9 different MR imaging modalities and labeled them based on relative CD3δ expression. The best architecture we were able to devise is based on ResNet50, using an Adam optimizer and binary cross-entropy loss function. The result is a model with a test set binary accuracy of 0.5667 and AUC of 0.5136. This task is impossible for trained radiologists so this shows some promise.

Hyperparamenter tuning had minimal effect of our success, which suggests that feature engineering (including data augmentation) and model selection needs significant work. With more time, team members, and computational resources, we would clean up the feature images by removing incomplete examples. We could try enlarging the individual features by including multiple z-slices (tying the voxels to physical dimensions) and resizing using interpolation. We could augment our data using rotation and flipping. We could also experiment with labels other than CD3δ to see if there are better descriptors of our features.

Once we have a satisfactory dataset, a larger ResNet or even DenseNet architecture may give us better results. We can look to see if there are pre-trained parameters that would be useful for transfer learning despite our quite unique application.

Ultimately, we want to deploy a high-performing model to generate pheno-region maps of immune cell molecular signatures superimposed on 3D brain MR images. Achieving such a goal will facilitate clinical trials of immunotherapies for brain cancer and offer hope to many patients.

7 Contributions

Dileep Monie was the sole author of this report and the associated code (except where cited). Dr. Kyle Singleton managed the JupyterLab Docker environment used for this project. CS230 TA Shubhang Desai along with Drs. Andrea Hawkins-Daarud, Sara Ranjbar, Lee Curtin, and Pamela Jackson provided guidance and mentorship. Dr. Kristin Swanson provided access to the data and computational resources.

References

[1] Szabo, P.A., Levitin, H.M., Miron, M. et al. Single-cell transcriptomics of human T cells reveals tissue and activation signatures in health and disease. *Nat Commun* 10, 4706 (2019).

[2] Hu, L.S. et al. Radiogenomics to characterize regional genetic heterogeneity in glioblastoma. *Neuro Oncol* 19, 128–137 (2017).

[3] Hu, L.S. et al. Multi-Parametric MRI and Texture Analysis to Visualize Spatial Histologic Heterogeneity and Tumor Extent in Glioblastoma. *PLoS One* 10, e0141506 (2015).

[4] Hu, L.S. et al. Accurate patient-specific machine learning models of glioblastoma invasion using transfer learning. *Neuro Oncol* 19, vi157–vi158 (2017).

[5] Swanson, K.R. et al. Radiomics of tumor invasion 2.0: combining mechanistic tumor invasion models with machine learning models to accurately predict tumor invasion in human glioblastoma patients. *Neuro Oncol* 19, vi159–vi159 (2017).

[6] Haralick, R.M., Shanmugam, K., & Dinstein, I.H. Textural features for image classification. *IEEE Transactions on systems, man, and cybernetics*, (6), 610-621 (1973).

[7] Feichtinger, H.G., & Strohmer, T. (Eds.). Gabor analysis and algorithms: Theory and applications. *Springer Science & Business Media*, (2012).

[8] Hu, L.S. et al. Multi-parametric MRI and texture analysis to visualize spatial histologic heterogeneity and tumor extent in glioblastoma. *PloS One*, 10(11), e0141506 (2015).

[9] Yang, W., Wang, K., & Zuo, W. Neighborhood Component Feature Selection for High-Dimensional Data. *JCP*, 7(1), 161-168 (2012).

[10] Hawkins-Daarud, A. et al. Revealing the tumor-immune landscape through spatially-resolved radiomics: case studies. *Neuro Oncol* 21, vi169–vi170 (2019).

[11] Wang, J.H. et al. Radiomic biomarkers of tumor immune biology and immunotherapy response. *Clin Trans Rad Oncol*, 28, 97-115 (2021).

[12] He, K., Zhang, X., Ren, S., & Sun, J. Deep Residual Learning for Image Recognition. arXiv:1512.03385v1 [cs.CV] (2015).

[13] Kingma, D.P. & Ma, J. Adam: A Method for Stochastic Optimization. arXiv:1412.6980v9 [cs.LG] (2017).

[14] Huang, G., Liu, Z., van der Maaten, L., & Weinberger, K.Q. Densely Connected Convolutional Networks. arXiv:1608.06993v5 [cs.CV] (2018).