Project Report

Using Deep Learning To Classify Sub-Visual Glioblastoma Infiltration Of The Brainstem

Abdullahi Elmi

Supervisor: Aggelos K. Katsaggelos

Abstract:

Background & Purpose:

Current identification of glioblastoma infiltration of the brainstem relies on a visual assessment of MRI imaging by a medical professional, which can be prone to a lot of error in the early stages of identification due to the microscopic nature of the infiltration at first. This project aims to identify if a deep-learning model has the potential to classify the presence of glioblastoma infiltration from MRI alone.

Materials & Methods:

There were 4 MRI pulse sequences (T1, T1c, T2, FLAIR) taken for each patient, with a total of 33 patients collected, 3 of whom had missing MRI sequences, 6 of whom though had missing data entirely and were excluded from the model, leaving a final total of 24. Densely Connected Convolutional Networks (DenseNet) was the model utilized for classification, with binary cross entropy being used to calculate loss, and early stopping also implemented to prevent overfitting. A confusion matrix was computed for a tuned model on the entire dataset, and Leave One Out Cross Validation (LOOCV) was used to get a more accurate understanding of the model's performance.

Results:

With the limited dataset, the model at first glance was able to avoid any severe overfitting, with the validation loss reaching a minimum of ~0.5431 while the training loss was at around ~0.36 at that minimum validation step (and had an overall minimum of ~0.21). The model, when run on the entire dataset, was able to achieve an accuracy of 0.75, a sensitivity of 0.73 & a specificity of 0.78. When fully implementing LOOCV, the model failed entirely to properly classify negative cases, achieving an accuracy of 0.625, sensitivity of 1.00, & specificity of 0.00.

Conclusion:

This project has indicated that there are not some easy to distinguish features in sub-visual infiltration by glioblastoma that a deep learning model could pick up on without guidance. That being said, it simply rules out the simplest/most convenient option; and with a more robust dataset (especially in its control group) and with well-informed feature selection, it may still be possible for a model to classify infiltration with conventional MRI alone.

Background:

Glioblastomas are a grade IV glioma (the highest grade), and the most aggressive type of brain tumor found in adults, with an incidence of 3.19 per 100,000 population [1] and median survival length of ~15 months [2]. Currently, diagnosis and prognosis rely on medical professionals analyzing conventional Magnetic Resonance Imaging (MRI), including pulse sequences such as T1 weighted contrast-enhanced (ce) and T2 weighted/FLAIR (Fluid attenuated inversion recovery) MRI. This method, while sufficient in identifying the presence of tumors, is not so when identifying low-density tumor cell infiltration beyond the contrast enhancing region, which is commonly occurring for high-grade gliomas [3]. Recognizing the infiltration of these cancer cells would help oncologists give patients more accurate life expectancies, recognize tumor recurrence earlier, and possibly extend their life expectancy by increasing the treatment efficacy, however limited it may be for glioblastomas.

These pitfalls in analyzing conventional MRI for infiltration in high grade gliomas like glioblastoma, has lead some researchers to utilize multimodal imaging, by combining conventional MRI with advanced modalities such as diffusion tensor imaging (DTI) and dynamic susceptibility contrast (DSC), and then implementing AI for the delineation of diffuse glioma infiltration [4, 5, 6, 7]. These papers have had promising results, but as much research has not been made in attempting to classify or delineate glioblastoma infiltration with conventional MRI alone. The aim of this project is to try and classify glioblastoma infiltration using only conventional MRI.

Materials and Methods:

MRI data was collected from a total of 33 patients, but due to missing data or MRI sequences, there were only 24 patients with full data and MRI sequences. Every patient is meant to have 4 MRI pulse sequences, T1, T1c, T2, & FLAIR (Fluid Attenuated Inversion Recovery), as well as a final segmentation image that would highlight the brainstem. Initially the patients were classified into 3 categories of having no infiltration, microscopic infiltration, and extensive infiltration. The dataset was incredibly unbalanced with only two patients having no infiltration at all, so the no-infiltration and microscopic categories were combined, effectively focusing the model on attempting to identify the presence of extensive infiltration. There were 15 extensive infiltration patients of the final 24, and 9 microscopic infiltration / no infiltration patients.

For handling the data and model, a number of python libraries/frameworks were used, specifically, TorchIO, MONAI, & PyTorch Lightning, all of which were compatible with and built as an extension of PyTorch. A number of preprocessing operations were applied to the images: Reordering the data to be closest to canonical (RAS+) orientation, resampling the different MRI sequences to the same physical space, rescaling the intensity, cropping to isolate mainly the brainstem, and reencoding label maps with one hot encoding. Augmentation operations were also applied to training images, namely: random affine transformations, adding random gaussian noise, randomly changing contrast by raising their values to the power of a gamma parameter, simulating random MRI motion artifacts, & random MRI bias field artifacts.

The data was then split into training and validation sets of 19 & 5 respectively, and a batch size of 4 had been used in training, and a size of 5 for validating. The model that was used was Densely Connected Convolutional Networks (DenseNet), specifically DenseNet121, with Binary Cross Entropy being used to log training and validation losses. The validation loss was also tracked as a metric to implement checkpointing and early stopping, where after 3 (2 during LOOCV) epochs without improvement, training of the model would cease.

Learning rates were explored with trial and error, beginning with 1e^{-2} that caused an exploding gradient, and led to NaN validation losses, while learning rates of 1e^{-6} and below didn't learn fast enough. A learning rate of ~ 1e^{-3} - 1e^{-4} seemed ideal at first, but 1e^{-3} was too volatile, it would start with an incredibly high validation loss in the dozens to hundreds but could drop down to a low of ~0.6672, garnering good results at times, but not reliably. A learning rate of 1e^{-4} would start with a much lower validation loss of around ~1.5, but would also not improve as well, reaching minimum validation losses of ~0.7609.

Lastly, Leave One Out Cross Validation (LOOCV) was also used in order to better gauge the performance of the model considering what little data is present. With 24 MRI sets in total, and training 24 separate models (each being tested on only one unique left out MRI set), the LOOCV had to be run across 4 separate Colab notebooks/Google Accounts due to GPU limitations, although they're identical save for the ranges of indices looped over, so the one example notebook linked below should be representative.

Results:

The final learning rate that was landed upon was 3e^{-4}, as it achieved a minimum validation loss of 0.5431 (See Figures 1 & 2), while the training loss at that same step was at ~0.36 and reached a minimum of ~0.21 during that training session. On repeated attempts it also had reliable results. With that learning rate, a batch size of 4 for training and 5 for validating, the trained model when applied to the entire dataset (training + validation), would achieve an accuracy of 0.75, a sensitivity of 0.73 & a specificity of 0.78 (See Figure 3).

LOOCV was also performed on the dataset, and results were not as one would have hoped. The models when performing a single prediction leaned entirely towards the positive classification. So LOOCV had an overall accuracy of 0.625, sensitivity of 1.00, & specificity of 0.00.

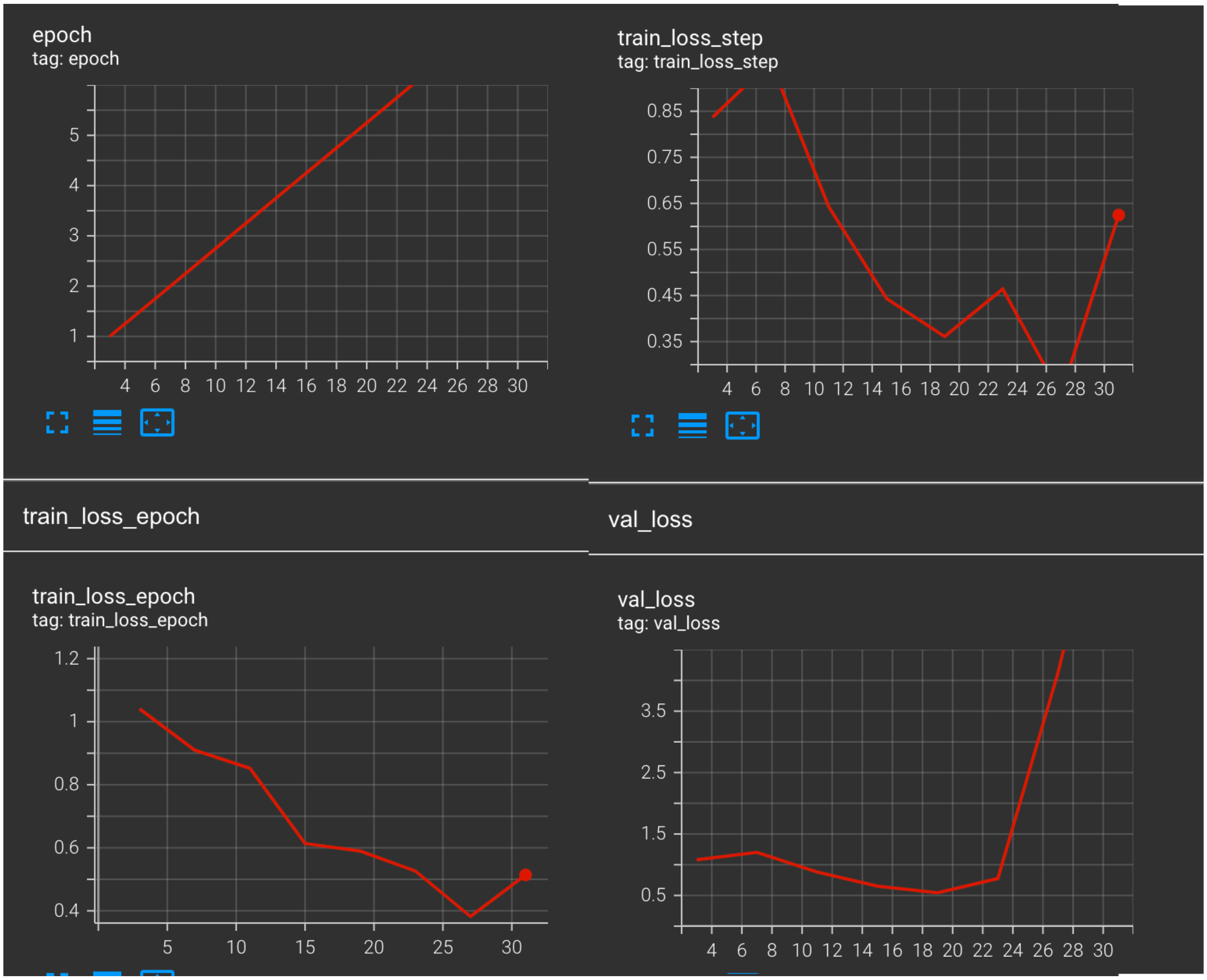
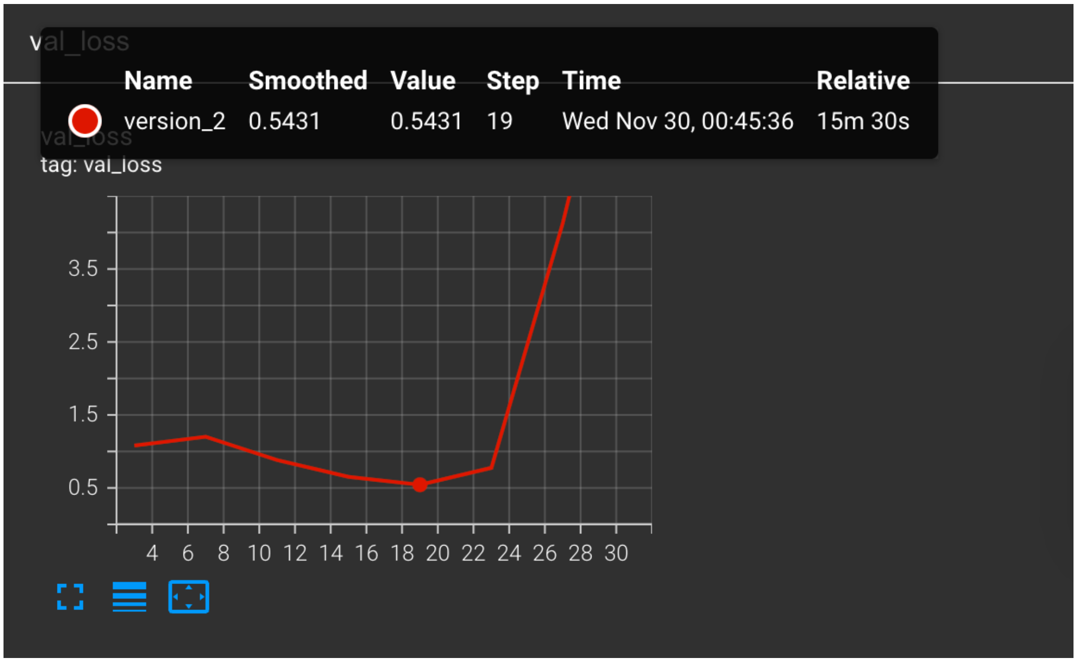
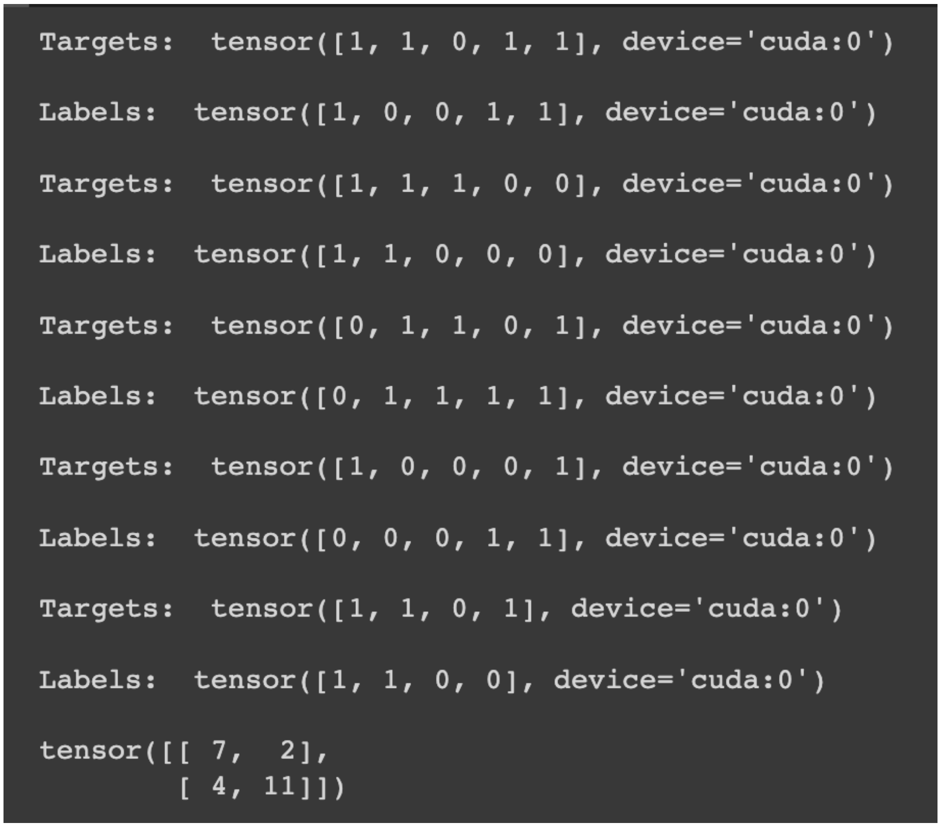


Figure 1: Training & Validation Loss Example For lr=3e^{-4}



Name	Smoothed	Value	Step	Time	Relative
version_2	0.5431	0.5431	19	Wed Nov 30, 00:45:36	15m 30s

Figure 2: Same Val_Loss graph as in Figure 1, simply highlighting the minimum val_loss



```
Targets:  tensor([1, 1, 0, 1, 1], device='cuda:0')

Labels:  tensor([1, 0, 0, 1, 1], device='cuda:0')

Targets:  tensor([1, 1, 1, 0, 0], device='cuda:0')

Labels:  tensor([1, 1, 0, 0, 0], device='cuda:0')

Targets:  tensor([0, 1, 1, 0, 1], device='cuda:0')

Labels:  tensor([0, 1, 1, 1, 1], device='cuda:0')

Targets:  tensor([1, 0, 0, 0, 1], device='cuda:0')

Labels:  tensor([0, 0, 0, 1, 1], device='cuda:0')

Targets:  tensor([1, 1, 0, 1], device='cuda:0')

Labels:  tensor([1, 1, 0, 0], device='cuda:0')

tensor([[ 7,  2],
        [ 4, 11]])
```

Figure 3: Running Trained Model on entire dataset (training + validation). Format: [[TN, FP], [FN, TP]]

Discussion:

Initially, training the model and simultaneously working to get its validation loss lower with a multitude of learning rates would indicate that the DenseNet was capable of promising results without severe overfitting due to the converging training and validation losses & sufficient predictions that managed to have encouraging specificity considering the imbalance in the data. The combination of a significantly decreasing validation loss & equivalent specificity and sensitivity when running the model on the entirety of the data after training on it led to the belief that while it was almost certainly overtrained to an extent, it was capable of some generalization.

It was prudent then that LOOCV was properly implemented as it's shown that the model was not at all sufficiently capable of predicting negative cases that it has not seen during training. While disheartening, it was still an outcome to have expected. The imbalance of the dataset has certainly played a major part in this, and it may even be possible that lumping microscopic infiltration in with the negative cases exasperated this, as there could be features shared between extensive and microscopic infiltration cases that the model has internalized, although the two entirely healthy patients still being misclassified points to it being more than that. It seems certain now that there aren't differences a model can pick up on its own unguided, and that feature selection would likely be a requirement for there to be a chance of having a deep learning model classify sub-visual infiltration by glioblastomas, as the alternative would be having massive swathes of data to essentially brute-force it, which isn't particularly realistic when discussing medical data.

This may be a motivating factor for why much of the research on using machine learning to delineate glioma infiltration relies on multimodal imaging (beyond just MRI), as most of them, like in [4, 5, 6, 7], also suffer from small datasets, but are provided with far more in-depth information per patient via multiple modalities. That being said, some of these studies don't use separate validation sets or cross-validation to evaluate their model's performance [7], so it may still be a significant limiting factor in their case as well.

Conclusion:

This project has not succeeded in classifying sub-visual glioblastoma infiltration of the brainstem with machine learning and conventional MRI alone. It has essentially ruled out the option of achieving it without robust data/feature selection, and so these improvements could still be worth pursuing in a subsequent project. Beyond that, common image modalities such as DTI are not prohibitively expensive, and success has already been achieved on that front in not only classifying but delineating infiltration, so further research efforts may be better spent on taking full advantage of multimodal imaging. While utilizing MRI alone to classify sub-visual infiltration would be incredibly convenient, the prospect of identifying the kinds of sub-visual infiltration neurologists can't see with the naked eye with multimodal imaging would still achieve the goal of better equipping medical professionals with the tools to give their patients the best possible care.

Google Drive Links:

Model Notebook:

- https://colab.research.google.com/drive/1-i-tuFoEx8bYcaP_aHx7NTjnI5-9xNv9?usp=sharing

LOOCV Notebook (pretty similar to the first, just implementing the core loop for LOOCV):

- https://colab.research.google.com/drive/1bmaXe-N6yvtS2BntkIkUm-PjEHMqkadD?usp=sharing

Dataset Zip File to run it yourself:

- https://drive.google.com/file/d/1XlHWz0O_x9ZDn1JDFuEdMPjh6Q_SNeKQ/view?usp=sharing

Example Model:

- https://drive.google.com/file/d/1NtVdwvmYKl0WAz8VhxllFzK7WTFZcemb/view?usp=sharing
- If you don't wanna train a model yourself, this is an example model that yielded the confusion matrix in Figure 3 (the results when applying the model to all of the data, not a LOOCV one)
- For instructions: Add it to your colab runtime files for the Model Notebook, copy the path, and paste it under the commented line "# Loading A Model From /content".

Citations:

1. Tamimi AF, Juweid M. Epidemiology and Outcome of Glioblastoma. In: De Vleeschouwer S, editor. Glioblastoma [Internet]. Brisbane (AU): Codon Publications; 2017 Sep 27. Chapter 8. Available from: https://www.ncbi.nlm.nih.gov/books/NBK470003/ doi: 10.15586/codon.glioblastoma.2017.ch8
2. Stupp, R., Mason, W. P., Van Den Bent, M. J., Weller, M., Fisher, B., Taphoorn, M. J., ... & Mirimanoff, R. O. (2005). Radiotherapy plus concomitant and adjuvant temozolomide for glioblastoma. *New England journal of medicine*, *352*(10), 987-996.
3. Eidel, O., Burth, S., Neumann, J. O., Kieslich, P. J., Sahm, F., Jungk, C., Kickingereder, P., Bickelhaupt, S., Mundiyanapurath, S., Bäumer, P., Wick, W., Schlemmer, H. P., Kiening, K., Unterberg, A., Bendszus, M., & Radbruch, A. (2017). Tumor Infiltration in

Enhancing and Non-Enhancing Parts of Glioblastoma: A Correlation with Histopathology. *PloS one*, *12*(1), e0169292. https://doi.org/10.1371/journal.pone.0169292

4. Akbari, H., Macyszyn, L., Da, X., Bilello, M., Wolf, R. L., Martinez-Lage, M., Biros, G., Alonso-Basanta, M., O'Rourke, D. M., & Davatzikos, C. (2016). Imaging Surrogates of Infiltration Obtained Via Multiparametric Imaging Pattern Analysis Predict Subsequent Location of Recurrence of Glioblastoma. *Neurosurgery*, *78*(4), 572–580. https://doi.org/10.1227/NEU.0000000000001202
5. Rathore, S., Akbari, H., Doshi, J., Shukla, G., Rozycki, M., Bilello, M., Lustig, R., & Davatzikos, C. (2018). Radiomic signature of infiltration in peritumoral edema predicts subsequent recurrence in glioblastoma: implications for personalized radiotherapy planning. *Journal of medical imaging (Bellingham, Wash.)*, *5*(2), 021219. https://doi.org/10.1117/1.JMI.5.2.021219
6. Yan, J. L., Li, C., van der Hoorn, A., Boonzaier, N. R., Matys, T., & Price, S. J. (2020). A Neural Network Approach to Identify the Peritumoral Invasive Areas in Glioblastoma Patients by Using MR Radiomics. *Scientific reports*, *10*(1), 9748. https://doi.org/10.1038/s41598-020-66691-6
7. Lipkova, J., Angelikopoulos, P., Wu, S., Alberts, E., Wiestler, B., Diehl, C., Preibisch, C., Pyka, T., Combs, S. E., Hadjidoukas, P., Van Leemput, K., Koumoutsakos, P., Lowengrub, J., & Menze, B. (2019). Personalized Radiotherapy Design for Glioblastoma: Integrating Mathematical Tumor Models, Multimodal Scans, and Bayesian Inference. *IEEE transactions on medical imaging*, *38*(8), 1875–1884. https://doi.org/10.1109/TMI.2019.2902044