Tumor Recurrence Prediction After Stereotactic Radiosurgery: Multimodal Machine Learning Approaches in Imbalanced Data Contexts

Stiven Hidri, Santa Di Cataldo, Francesco Ponzio

November 2024

1 Summary

Brain tumors are abnormal growths of cells that can cause several health issues by applying pressure on critical brain structures. One of the main procedures to treat brain tumors is surgery, which aims to remove as much of the tumor mass as possible, hopefully helping ongoing treatments. Gamma Knife, on the other hand, allows to treat the tumor without performing any incision by focusing several radiation beams specifically on the lesion location. The radiation dose along each beam is low, but at the focal point is significantly higher. The outcome of such treatment cannot be assessed immediately after: follow-ups will determine how the lesion responds to the treatment. Being able to correctly predict the outcome of such procedure can improve treatment planning and design of patients affected by brain tumors. In this thesis we address the task of narrowing down the prediction to individual lesions, with the aim of further improving treatment personalization.

The dataset contains pre-treatment CET1W MRI scans, planned radiation dose maps, ROI masks and clinical records. The main objective for the realization of this dataset, as stated from the authors, is to promote the development of machine learning solutions addressing the task of local recurrence prediction after Gamma Knife treatment. Gathered information include 47 patients affected by metastatic brain tumors. Overall, are present 244 lesions with annotations: 221 are labeled as 'stable' while the remaining 23 are labeled as 'recurrent'. The former includes Complete Response (CR), Partial Response (PR) and No Change (NC) whilst the latter indicates Progressive Diseases (PD). Eventual Beside or Remote Recurrence are treated as new lesions. Data preprocessing starts with the imaging: initially MRI scans are resampled to a pixel spacing of (1,1,1). Then, follow along ROI masks and radiation dose maps that are resampled too, by taking as reference the resampled MRI scans. The final results consist of aligned imaging where each voxel represents $1mm^3$ of physical space.

Finally, we used the ROI masks to isolate each lesion from the whole MRI scan with the relative radiation dose map.

Next, we associated each lesion and relative radiation dose maps to the corresponding clinical records, which contain: label (stable or recurrent), lesion location, primary tumor location, tumor histology, age at diagnosis, gender, fractions and months elapsed from the treatment to the follow-up. The categorical values did not follow a uniform and standardized nomenclature. Not only the values were inhomogeneous, but they also contained different levels of specificity, which led to a non trivial encoding process. To address these issues, string manipulation was performed to standardize all the terms. Then, in order to lower the number of features as a result of one hot encoding, we generalized the lesion locations by mapping each value to the corresponding brain region (parietal, frontal, occipital, temporal and cerebellum), along with its side (left or right). If the lesion location was not part of any aforementioned region, the original value was left as it was. Finally, categorical clinical values were one hot encoded. Normalization was performed through min-max scaling for both imaging and non-categorical clinical data.

We have chosen focal loss to address class imbalance: it allows the model to focus on misclassified samples by adding a regulating factor to the cross entropy loss. To improve generalization, employed data augmentation techniques were random flip, random rotation and random affine transformation. Random flip is applied along one of the axis with probability 0.5. Random rotation consist of randomly rotating the imaging with an angle between 0, 90, 180, 270 degrees along one of the axis. Lastly, random affine transformation rotates the targets with an angle between -30 and 30, scale with a factor between .85 and 1.15 and translates with an offset between -5 and 5.

The proposed deep learning models actuate different multi-modal fusion strategies. The BASE MODEL, proposed alongside the dataset, concatenates the features extracted from each modality through different deep learning backbones. DWT CONV, performs voxel-level input fusion by exploiting Discrete Wavelet Transforms. Then a 3D backbone, following the lines of the BASE MODEL, extracts relevant features from the fused input representation. Furthermore, the CONV LSTM network makes use of LSTM to fuse features extracted from overlapping 2D slices of both MRI and radiation dose maps. Finally, TRANS MED makes use of parallel self-attention mechanisms to find relevant relationships between features extracted from different modalities.

In order to make the most out of the limited samples, 10-fold cross validation has been applied on the whole dataset: 9 folds were used for training and 1 for testing. From the training set, an enough number of samples was held out to constitute a validation set. Then, to perform additional experiments and have a comparison with the baseline evaluations disclosed by the dataset's article, we maintained the same test set and applied 5-fold cross validation on the remaining training set: 4 folds were used for training and 1 for validation.

The results obtained by cross-validating the whole datasets were minimal, but not satisfactory. The best performing models, for what concerns the Area Under the Precision Recall Curve (AUPR), were BASE MODEL and WDT

CONV that reached respectively a score of 0.17 and 0.18 with an Area Under the Receiver Operating Characteristic curve (AUROC) of 0.67 and 0.6. Next, while keeping the fixed test set, we recorded the average performances along with minimums and maximums. Interestingly enough, we have noticed a significant difference between minimum and maximum metrics which can be interpreted as a sign of lack of generalization whenever the degree of difference of the sample distribution between sets is too high. The higher average AUPR was reached by the WDT CONV model: 0.313 with a corresponding averaged AUROC of 0.725. We also used a straightforward strategy to make use of the knowledge gathered from the models trained on different cross-validation splits: by averaging the prediction WDT CONV reached the highest AUPR value of 0.39 with a corresponding AUROC of 0.83, which shows the potential of combined effort between models trained on different splits. A final consideration to be made is that the chosen threshold was not always optimal, which can be clearly visible by considering the minimum values of precision and recall that often reached 0. This can be considered again as a sign of high degree of difference of sample distribution between sets.

The dataset used in this thesis presented many obstacles: both the data imbalance and the scarcity of the minority class undermined the models performances. Moreover, since this task focuses on individual lesion, smaller ones needed a more aggressive padding with respect to the bigger ones in order to reach the same input shape. This introduces artificial edges, that the models are going to learn anyway. Moreover, smaller lesions do contain very few voxels, which can lead to poor feature extractions. Surely, a higher spatial resolution might help to uncover valuable structural information of smaller tumors. Additionally, related works showed the high importance of complementary imaging which delivers valuable insight on how the tumor is affecting surrounding tissues. Furthermore, several works already performed exhaustive searches determining which clinical features are the most discriminant for this task (GPA score, KPS index, number of lesions). While others can be further studied, future datasets should preemptively take them into considerations. Finally, it is crucial to adopt standardized formatting for categorical values in order to ensure consistency and free selection of specificity for maximum exploitation of such features.

We hope that the results obtained in this work along with the produced considerations will give a solid help for future works addressing the prediction of local tumor recurrence after Gamma Knife therapy.